

# Introduction to computing with finite difference methods

Hans Petter Langtangen<sup>1,2</sup>

<sup>1</sup>Center for Biomedical Computing, Simula Research Laboratory

<sup>2</sup>Department of Informatics, University of Oslo

Apr 25, 2015

## Contents

<b>1</b>	<b>Finite difference methods</b>	<b>5</b>
1.1	A basic model for exponential decay . . . . .	5
1.2	The Forward Euler scheme . . . . .	6
1.3	The Backward Euler scheme . . . . .	10
1.4	The Crank-Nicolson scheme . . . . .	10
1.5	The unifying $\theta$ -rule . . . . .	12
1.6	Constant time step . . . . .	13
1.7	Mathematical derivation of finite difference formulas . . . . .	14
1.8	Compact operator notation for finite differences . . . . .	16
<b>2</b>	<b>Implementation</b>	<b>18</b>
2.1	Making a solver function . . . . .	20
2.2	Verifying the implementation . . . . .	25
2.3	Computing the numerical error as a mesh function . . . . .	27
2.4	Computing the norm of the numerical error . . . . .	28
2.5	Experiments with computing and plotting . . . . .	31
2.6	Memory-saving implementation . . . . .	35
<b>3</b>	<b>Exercises</b>	<b>37</b>
<b>4</b>	<b>Analysis of finite difference equations</b>	<b>39</b>
4.1	Experimental investigation of oscillatory solutions . . . . .	41
4.2	Exact numerical solution . . . . .	44
4.3	Stability . . . . .	45
4.4	Comparing amplification factors . . . . .	46
4.5	Series expansion of amplification factors . . . . .	46

4.6	The fraction of numerical and exact amplification factors . . . .	49
4.7	The global error at a point . . . . .	49
4.8	Integrated errors . . . . .	50
4.9	Truncation error . . . . .	51
4.10	Consistency, stability, and convergence . . . . .	52
<b>5</b>	<b>Exercises</b>	<b>54</b>
<b>6</b>	<b>Model extensions</b>	<b>55</b>
6.1	Generalization: including a variable coefficient . . . . .	55
6.2	Generalization: including a source term . . . . .	56
6.3	Implementation of the generalized model problem . . . . .	56
6.4	Verifying a constant solution . . . . .	58
6.5	Verification via manufactured solutions . . . . .	59
6.6	Computing convergence rates . . . . .	61
6.7	Extension to systems of ODEs . . . . .	62
<b>7</b>	<b>General first-order ODEs</b>	<b>63</b>
7.1	Generic form of first-order ODEs . . . . .	63
7.2	The $\theta$ -rule . . . . .	64
7.3	An implicit 2-step backward scheme . . . . .	64
7.4	Leapfrog schemes . . . . .	65
7.5	The 2nd-order Runge-Kutta method . . . . .	65
7.6	A 2nd-order Taylor-series method . . . . .	66
7.7	The 2nd- and 3rd-order Adams-Bashforth schemes . . . . .	66
7.8	The 4th-order Runge-Kutta method . . . . .	67
7.9	The Odespy software . . . . .	68
7.10	Example: Runge-Kutta methods . . . . .	69
7.11	Example: Adaptive Runge-Kutta methods . . . . .	71
<b>8</b>	<b>Exercises</b>	<b>73</b>
<b>9</b>	<b>Applications of exponential decay models</b>	<b>77</b>
9.1	Scaling . . . . .	77
9.2	Evolution of a population . . . . .	78
9.3	Compound interest and inflation . . . . .	79
9.4	Radioactive decay . . . . .	80
9.5	Newton's law of cooling . . . . .	82
9.6	Decay of atmospheric pressure with altitude . . . . .	83
9.7	Compaction of sediments . . . . .	84
9.8	Vertical motion of a body in a viscous fluid . . . . .	85
9.9	Decay ODEs from solving a PDE by Fourier expansions . . . . .	89
<b>10</b>	<b>Exercises</b>	<b>90</b>
<b>11</b>	<b>Summarizing multiple-choice questions</b>	<b>99</b>

## List of Exercises, Problems, and Projects

Exercise	1	Differentiate a function	p. 37
Exercise	2	Experiment with integer division	p. 38
Exercise	3	Experiment with wrong computations	p. 38
Exercise	4	Plot the error function	p. 38
Exercise	5	Change formatting of numbers and debug	p. 38
Exercise	6	Visualize the accuracy of finite differences	p. 54
Exercise	7	Explore the $\theta$ -rule for exponential ...	p. 54
Exercise	8	Experiment with precision in tests and the ...	p. 73
Exercise	9	Implement the 2-step backward scheme	p. 74
Exercise	10	Implement the 2nd-order Adams-Bashforth scheme ...	p. 74
Exercise	11	Implement the 3rd-order Adams-Bashforth scheme ...	p. 74
Exercise	12	Analyze explicit 2nd-order methods	p. 74
Problem	13	Implement and investigate the Leapfrog scheme	p. 74
Problem	14	Make a unified implementation of many schemes	p. 76
Exercise	15	Derive schemes for Newton's law of cooling	p. 90
Exercise	16	Implement schemes for Newton's law of cooling	p. 91
Exercise	17	Find time of murder from body temperature	p. 92
Exercise	18	Simulate an oscillating cooling process	p. 92
Exercise	19	Radioactive decay of Carbon-14	p. 92
Exercise	20	Simulate stochastic radioactive decay	p. 93
Exercise	21	Radioactive decay of two substances	p. 93
Exercise	22	Simulate the pressure drop in the atmosphere	p. 94
Exercise	23	Make a program for vertical motion in a fluid	p. 94
Project	24	Simulate parachuting	p. 95
Exercise	25	Formulate vertical motion in the atmosphere	p. 96
Exercise	26	Simulate vertical motion in the atmosphere	p. 97
Exercise	27	Compute $y =  x $ by solving an ODE	p. 97
Exercise	28	Simulate growth of a fortune with random interest ...	p. 97
Exercise	29	Simulate a population in a changing environment ...	p. 98
Exercise	30	Simulate logistic growth	p. 99
Exercise	31	Rederive the equation for continuous compound ...	p. 99
Exercise	32	Characterize a finite difference	p. 99
Exercise	33	Characterize a finite difference	p. 100
Exercise	34	The $\theta$ rule	p. 101
Exercise	35	What is the problem with this program?	p. 101
Exercise	36	Is the solution correct?	p. 101
Exercise	37	Is this a proper test function?	p. 102
Exercise	38	Recognize a programming language	p. 103
Exercise	39	Recognize a programming language	p. 103
Exercise	40	Recognize a programming language	p. 103
Exercise	41	Recognize a programming language	p. 104
Exercise	42	What is SymPy?	p. 104
Exercise	43	What kind of scheme is this?	p. 105
Exercise	44	What kind of scheme is this?	p. 105

Finite difference methods for partial differential equations (PDEs) employ a range of concepts and tools that can be introduced and illustrated in the context of simple ordinary differential equation (ODE) examples. This is what we do in the present document. By first working with ODEs, we keep the mathematical problems to be solved as simple as possible (but no simpler), thereby allowing full focus on understanding the key concepts and tools. The choice of topics in the forthcoming treatment of ODEs is therefore solely dominated by what carries over to numerical methods for PDEs.

Theory and practice are primarily illustrated by solving the very simple ODE  $u' = -au$ ,  $u(0) = I$ , where  $a > 0$  is a constant, but we also address the generalized problem  $u' = -a(t)u + b(t)$  and the nonlinear problem  $u' = f(u, t)$ . The following topics are introduced:

- How to think when constructing finite difference methods, with special focus on the Forward Euler, Backward Euler, and Crank-Nicolson (midpoint) schemes
- How to formulate a computational algorithm and translate it into Python code
- How to make curve plots of the solutions
- How to compute numerical errors
- How to compute convergence rates
- How to verify an implementation and automate verification through *test functions* in Python
- How to work with Python concepts such as arrays, lists, dictionaries, lambda functions, functions in functions (closures)
- How to perform array computing and understand the difference from scalar computing
- How to uncover numerical artifacts in the computed solution
- How to analyze the numerical schemes mathematically to understand why artifacts occur
- How to derive mathematical expressions for various measures of the error in numerical methods, frequently by using the `sympy` software for symbolic computation
- Introduce concepts such as finite difference operators, mesh (grid), mesh functions, stability, truncation error, consistency, and convergence
- Present additional methods for the general nonlinear ODE  $u' = f(u, t)$ , which is either a scalar ODE or a system of ODEs
- How to access professional packages for solving ODEs

- How the model equation  $u' = -au$  arises in a wide range of phenomena in physics, biology, and finance

#### The exposition in a nutshell.

Everything we cover is put into a practical, hands-on context. All mathematics is translated into working computing codes, and all the mathematical theory of finite difference methods presented here is motivated from a strong need to understand strange behavior of programs. Two fundamental questions saturate the text:

- How to we solve a differential equation problem and produce numbers?
- How to we trust the answer?

## 1 Finite difference methods

#### Goal.

We explain the basic ideas of finite difference methods using a simple ordinary differential equation  $u' = -au$  as primary example. Emphasis is put on the reasoning when discretizing the problem and introduction of key concepts such as mesh, mesh function, finite difference approximations, averaging in a mesh, derivation of algorithms, and discrete operator notation.

### 1.1 A basic model for exponential decay

Our model problem is perhaps the simplest ordinary differential equation (ODE):

$$u'(t) = -au(t),$$

Here,  $a > 0$  is a constant and  $u'(t)$  means differentiation with respect to time  $t$ . This type of equation arises in a number of widely different phenomena where some quantity  $u$  undergoes exponential reduction. Examples include radioactive decay, population decay, investment decay, cooling of an object, pressure decay in the atmosphere, and retarded motion in fluids (for some of these models,  $a$  can be negative as well), see Section 9 for details and motivation. We have chosen this particular ODE not only because its applications are relevant, but even more because studying numerical solution methods for this simple ODE gives important insight that can be reused in much more complicated settings, in particular when solving diffusion-type partial differential equations.

The analytical solution of the ODE is found by the method of separation of variables, which results in

$$u(t) = Ce^{-at},$$

for any arbitrary constant  $C$ . To formulate a mathematical problem for which there is a unique solution, we need a condition to fix the value of  $C$ . This condition is known as the *initial condition* and stated as  $u(0) = I$ . That is, we know the value  $I$  of  $u$  when the process starts at  $t = 0$ . The exact solution is then  $u(t) = Ie^{-at}$ .

We seek the solution  $u(t)$  of the ODE for  $t \in (0, T]$ . The point  $t = 0$  is not included since we know  $u$  here and assume that the equation governs  $u$  for  $t > 0$ . The complete ODE problem then reads: find  $u(t)$  such that

$$u' = -au, \quad t \in (0, T], \quad u(0) = I. \quad (1)$$

This is known as a *continuous problem* because the parameter  $t$  varies continuously from 0 to  $T$ . For each  $t$  we have a corresponding  $u(t)$ . There are hence infinitely many values of  $t$  and  $u(t)$ . The purpose of a numerical method is to formulate a corresponding *discrete problem* whose solution is characterized by a finite number of values, which can be computed in a finite number of steps on a computer.

## 1.2 The Forward Euler scheme

Solving an ODE like (1) by a finite difference method consists of the following four steps:

1. discretizing the domain,
2. fulfilling the equation at discrete time points,
3. replacing derivatives by finite differences,
4. formulating a recursive algorithm.

**Step 1: Discretizing the domain.** The time domain  $[0, T]$  is represented by a finite number of  $N_t + 1$  points

$$0 = t_0 < t_1 < t_2 < \dots < t_{N_t-1} < t_{N_t} = T. \quad (2)$$

The collection of points  $t_0, t_1, \dots, t_{N_t}$  constitutes a *mesh* or *grid*. Often the mesh points will be uniformly spaced in the domain  $[0, T]$ , which means that the spacing  $t_{n+1} - t_n$  is the same for all  $n$ . This spacing is often denoted by  $\Delta t$ , in this case  $t_n = n\Delta t$ .

We seek the solution  $u$  at the mesh points:  $u(t_n)$ ,  $n = 1, 2, \dots, N_t$ . Note that  $u^0$  is already known as  $I$ . A notational short-form for  $u(t_n)$ , which will be used extensively, is  $u^n$ . More precisely, we let  $u^n$  be the *numerical approximation* to the

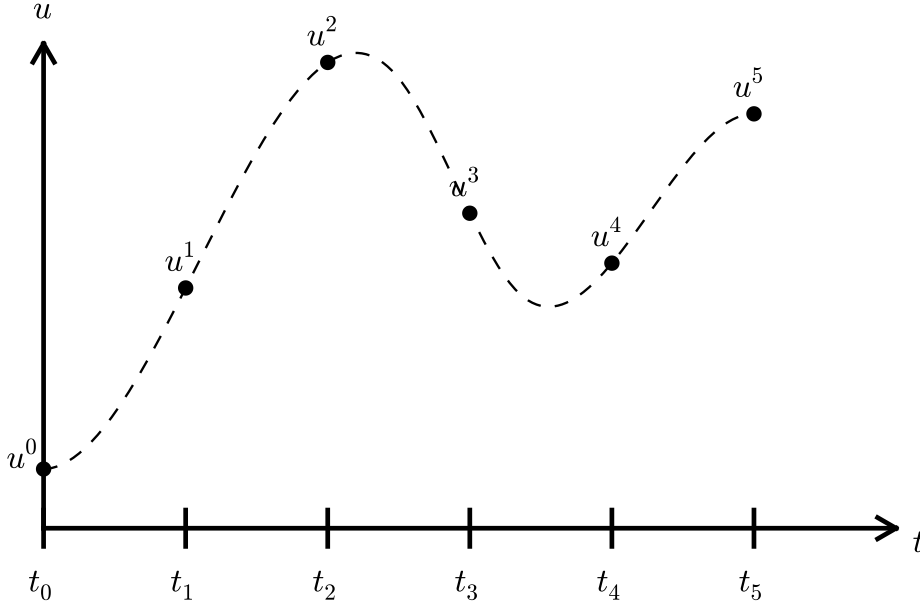


Figure 1: Time mesh with discrete solution values.

exact solution  $u(t_n)$  at  $t = t_n$ . The numerical approximation is a *mesh function*, here defined only at the mesh points. When we need to clearly distinguish between the numerical and the exact solution, we often place a subscript  $e$  on the exact solution, as in  $u_e(t_n)$ . Figure 1 shows the  $t_n$  and  $u_n$  points for  $n = 0, 1, \dots, N_t = 7$  as well as  $u_e(t)$  as the dashed line. The goal of a numerical method for ODEs is to compute the mesh function by solving a finite set of *algebraic equations* derived from the original ODE problem.

Since finite difference methods produce solutions at the mesh points only, it is an open question what the solution is between the mesh points. One can use methods for interpolation to compute the value of  $u$  between mesh points. The simplest (and most widely used) interpolation method is to assume that  $u$  varies linearly between the mesh points, see Figure 2. Given  $u^n$  and  $u^{n+1}$ , the value of  $u$  at some  $t \in [t_n, t_{n+1}]$  is by linear interpolation

$$u(t) \approx u^n + \frac{u^{n+1} - u^n}{t_{n+1} - t_n}(t - t_n). \quad (3)$$

**Step 2: Fulfilling the equation at discrete time points.** The ODE is supposed to hold for all  $t \in (0, T]$ , i.e., at an infinite number of points. Now we relax that requirement and require that the ODE is fulfilled at a finite set of discrete points in time. The mesh points  $t_0, t_1, \dots, t_{N_t}$  are a natural (but not the only) choice of points. The original ODE is then reduced to the following  $N_t$  equations:

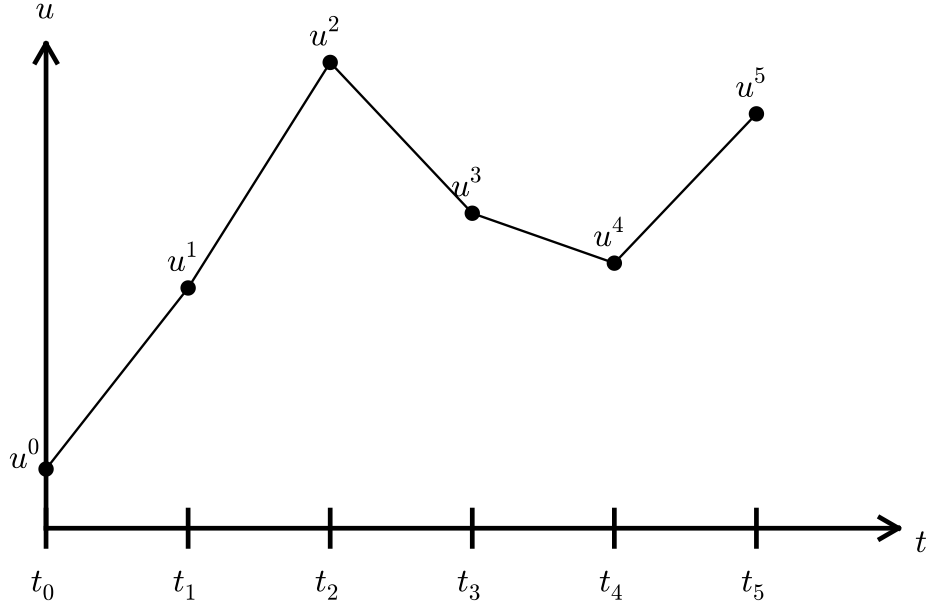


Figure 2: Linear interpolation between the discrete solution values (dashed curve is exact solution).

$$u'(t_n) = -au(t_n), \quad n = 0, \dots, N_t. \quad (4)$$

**Step 3: Replacing derivatives by finite differences.** The next and most essential step of the method is to replace the derivative  $u'$  by a finite difference approximation. Let us first try a one-sided difference approximation (see Figure 3),

$$u'(t_n) \approx \frac{u^{n+1} - u^n}{t_{n+1} - t_n}. \quad (5)$$

Inserting this approximation in (4) results in

$$\frac{u^{n+1} - u^n}{t_{n+1} - t_n} = -au^n, \quad n = 0, 1, \dots, N_t - 1. \quad (6)$$

Later it will be absolutely clear that if we want to compute the solution up to time level  $N_t$ , we only need (4) to hold for  $n = 0, \dots, N_t - 1$  since (6) for  $n = N_t - 1$  creates an equation for the final value  $u^{N_t}$ .

Equation (6) is the discrete counterpart to the original ODE problem (1), and often referred to as *finite difference scheme* or more generally as the *discrete equations* of the problem. The fundamental feature of these equations is that they are *algebraic* and can hence be straightforwardly solved to produce the mesh function, i.e., the values of  $u$  at the mesh points  $(u^n, n = 1, 2, \dots, N_t)$ .



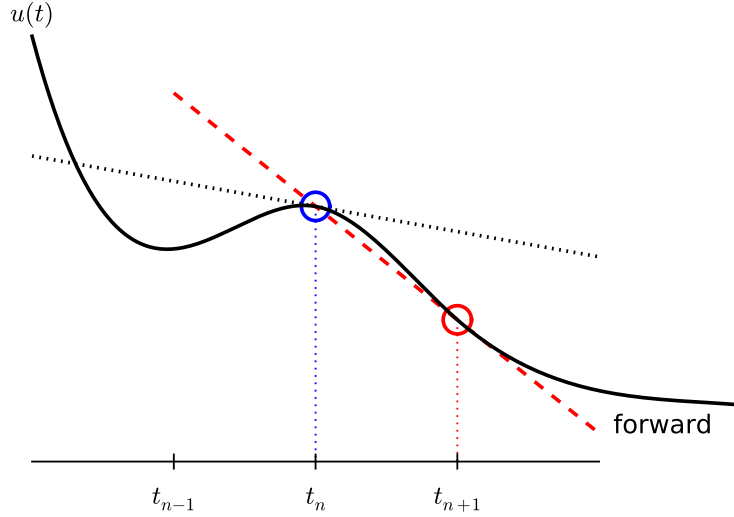


Figure 3: Illustration of a forward difference.

**Step 4: Formulating a recursive algorithm.** The final step is to identify the computational algorithm to be implemented in a program. The key observation here is to realize that (6) can be used to compute  $u^{n+1}$  if  $u^n$  is known. Starting with  $n = 0$ ,  $u^0$  is known since  $u^0 = u(0) = I$ , and (6) gives an equation for  $u^1$ . Knowing  $u^1$ ,  $u^2$  can be found from (6). In general,  $u^n$  in (6) can be assumed known, and then we can easily solve for the unknown  $u^{n+1}$ :

$$u^{n+1} = u^n - a(t_{n+1} - t_n)u^n. \quad (7)$$

We shall refer to (7) as the Forward Euler (FE) scheme for our model problem. From a mathematical point of view, equations of the form (7) are known as *difference equations* since they express how differences in  $u$ , like  $u^{n+1} - u^n$ , evolve with  $n$ . The finite difference method can be viewed as a method for turning a differential equation into a difference equation.

Computation with (7) is straightforward:

$$\begin{aligned} u_0 &= I, \\ u_1 &= u^0 - a(t_1 - t_0)u^0 = I(1 - a(t_1 - t_0)), \\ u_2 &= u^1 - a(t_2 - t_1)u^1 = I(1 - a(t_1 - t_0))(1 - a(t_2 - t_1)), \\ u^3 &= u^2 - a(t_3 - t_2)u^2 = I(1 - a(t_1 - t_0))(1 - a(t_2 - t_1))(1 - a(t_3 - t_2)), \end{aligned}$$

and so on until we reach  $u^{N_t}$ . Very often,  $t_{n+1} - t_n$  is constant for all  $n$ , so we can introduce the common symbol  $\Delta t$  for the time step:  $\Delta t = t_{n+1} - t_n$ ,  $n = 0, 1, \dots, N_t - 1$ . Using a constant time step  $\Delta t$  in the above calculations gives

$$\begin{aligned}
u_0 &= I, \\
u_1 &= I(1 - a\Delta t), \\
u_2 &= I(1 - a\Delta t)^2, \\
u_3 &= I(1 - a\Delta t)^3, \\
&\vdots \\
u^{N_t} &= I(1 - a\Delta t)^{N_t}.
\end{aligned}$$

This means that we have found a closed formula for  $u^n$ , and there is no need to let a computer generate the sequence  $u^1, u^2, u^3, \dots$ . However, finding such a formula for  $u^n$  is possible only for a few very simple problems, so in general finite difference equations must be solved on a computer.

As the next sections will show, the scheme (7) is just one out of many alternative finite difference (and other) methods for the model problem (1).

### 1.3 The Backward Euler scheme

There are several choices of difference approximations in step 3 of the finite difference method as presented in the previous section. Another alternative is

$$u'(t_n) \approx \frac{u^n - u^{n-1}}{t_n - t_{n-1}}. \quad (8)$$

Since this difference is based on going backward in time ( $t_{n-1}$ ) for information, it is known as the Backward Euler difference. Figure 4 explains the idea.

Inserting (8) in (4) yields the Backward Euler (BE) scheme:

$$\frac{u^n - u^{n-1}}{t_n - t_{n-1}} = -au^n. \quad (9)$$

We assume, as explained under step 4 in Section 1.2, that we have computed  $u^0, u^1, \dots, u^{n-1}$  such that (9) can be used to compute  $u^n$ . For direct similarity with the Forward Euler scheme (7) we replace  $n$  by  $n + 1$  in (9) and solve for the unknown value  $u^{n+1}$ :

$$u^{n+1} = \frac{1}{1 + a(t_{n+1} - t_n)} u^n. \quad (10)$$

### 1.4 The Crank-Nicolson scheme

The finite difference approximations used to derive the schemes (7) and (10) are both one-sided differences, known to be less accurate than central (or midpoint) differences. We shall now construct a central difference at  $t_{n+\frac{1}{2}} = \frac{1}{2}(t_n + t_{n+1})$ , or  $t_{n+\frac{1}{2}} = (n + \frac{1}{2})\Delta t$  if the mesh spacing is uniform in time. The approximation reads

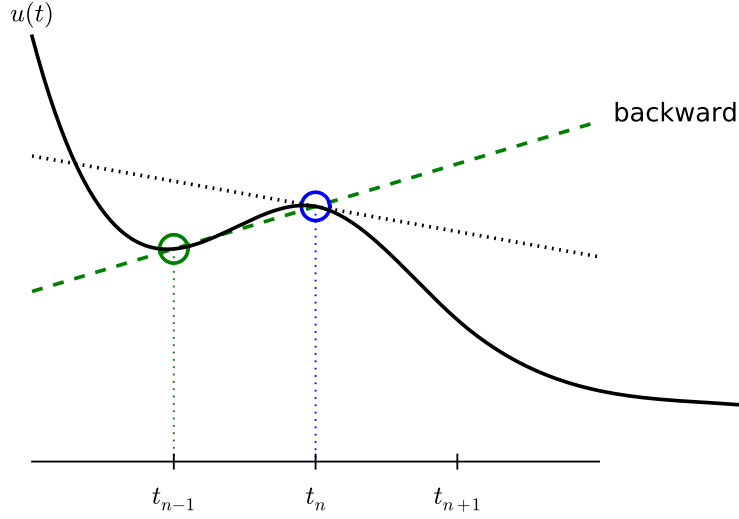


Figure 4: Illustration of a backward difference.

$$u'(t_{n+\frac{1}{2}}) \approx \frac{u^{n+1} - u^n}{t_{n+1} - t_n}. \quad (11)$$

Note that the fraction on the right-hand side is the same as for the Forward Euler approximation (5) and the Backward Euler approximation (8) (with  $n$  replaced by  $n + 1$ ). The accuracy of this fraction as an approximation to the derivative of  $u$  depends on *where* we seek the derivative: in the center of the interval  $[t_n, t_{n+1}]$  or at the end points.

With the formula (11), where  $u'$  is evaluated at  $t_{n+\frac{1}{2}}$ , it is natural to demand the ODE to be fulfilled at the time points *between* the mesh points:

$$u'(t_{n+\frac{1}{2}}) = -au(t_{n+\frac{1}{2}}), \quad n = 0, \dots, N_t - 1. \quad (12)$$

Using (11) in (12) results in

$$\frac{u^{n+1} - u^n}{t_{n+1} - t_n} = -au^{n+\frac{1}{2}}, \quad (13)$$

where  $u^{n+\frac{1}{2}}$  is a short form for  $u(t_{n+\frac{1}{2}})$ . The problem is that we aim to compute  $u^n$  for integer  $n$ , implying that  $u^{n+\frac{1}{2}}$  is not a quantity computed by our method. It must therefore be expressed by the quantities that we actually produce, i.e., the numerical solution at the mesh points. One possibility is to approximate  $u^{n+\frac{1}{2}}$  as an arithmetic mean of the  $u$  values at the neighboring mesh points:

$$u^{n+\frac{1}{2}} \approx \frac{1}{2}(u^n + u^{n+1}). \quad (14)$$

Using (14) in (13) results in

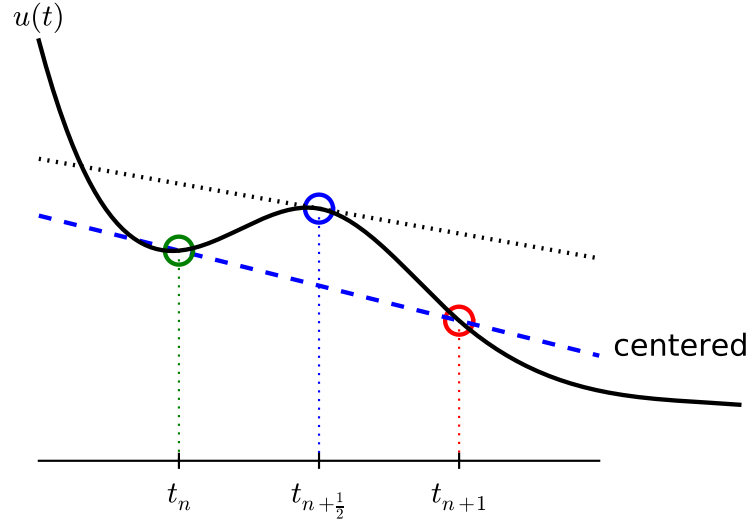


Figure 5: Illustration of a centered difference.

$$\frac{u^{n+1} - u^n}{t_{n+1} - t_n} = -a \frac{1}{2} (u^n + u^{n+1}). \quad (15)$$

Figure 5 sketches the geometric interpretation of such a centered difference.

We assume that  $u^n$  is already computed so that  $u^{n+1}$  is the unknown, which we can solve for:

$$u^{n+1} = \frac{1 - \frac{1}{2}a(t_{n+1} - t_n)}{1 + \frac{1}{2}a(t_{n+1} - t_n)} u^n. \quad (16)$$

The finite difference scheme (16) is often called the Crank-Nicolson (CN) scheme or a midpoint or centered scheme.

### 1.5 The unifying $\theta$ -rule

The Forward Euler, Backward Euler, and Crank-Nicolson schemes can be formulated as one scheme with a varying parameter  $\theta$ :

$$\frac{u^{n+1} - u^n}{t_{n+1} - t_n} = -a(\theta u^{n+1} + (1 - \theta)u^n). \quad (17)$$

Observe:

- $\theta = 0$  gives the Forward Euler scheme
- $\theta = 1$  gives the Backward Euler scheme, and
- $\theta = \frac{1}{2}$  gives the Crank-Nicolson scheme.

- We may alternatively choose any other value of  $\theta$  in  $[0, 1]$ .

As before,  $u^n$  is considered known and  $u^{n+1}$  unknown, so we solve for the latter:

$$u^{n+1} = \frac{1 - (1 - \theta)a(t_{n+1} - t_n)}{1 + \theta a(t_{n+1} - t_n)}. \quad (18)$$

This scheme is known as the  $\theta$ -rule, or alternatively written as the “theta-rule”.

#### Derivation.

We start with replacing  $u'$  by the fraction

$$\frac{u^{n+1} - u^n}{t_{n+1} - t_n},$$

in the Forward Euler, Backward Euler, and Crank-Nicolson schemes. Then we observe that the difference between the methods concerns which point this fraction approximates the derivative. Or in other words, at which point we sample the ODE. So far this has been the end points or the midpoint of  $[t_n, t_{n+1}]$ . However, we may choose any point  $\tilde{t} \in [t_n, t_{n+1}]$ . The difficulty is that evaluating the right-hand side  $-au$  at an arbitrary point faces the same problem as in Section 1.4: the point value must be expressed by the discrete  $u$  quantities that we compute by the scheme, i.e.,  $u^n$  and  $u^{n+1}$ . Following the averaging idea from Section 1.4, the value of  $u$  at an arbitrary point  $\tilde{t}$  can be calculated as a *weighted average*, which generalizes the arithmetic mean  $\frac{1}{2}u^n + \frac{1}{2}u^{n+1}$ . If we express  $\tilde{t}$  as a weighted average

$$t_{n+\theta} = \theta t_{n+1} + (1 - \theta)t_n,$$

where  $\theta \in [0, 1]$  is the weighting factor, we can write

$$u(\tilde{t}) = u(\theta t_{n+1} + (1 - \theta)t_n) \approx \theta u^{n+1} + (1 - \theta)u^n. \quad (19)$$

We can now let the ODE hold at the point  $\tilde{t} \in [t_n, t_{n+1}]$ , approximate  $u'$  by the fraction  $(u^{n+1} - u^n)/(t_{n+1} - t_n)$ , and approximate the right-hand side  $-au$  by the weighted average (19). The result is (17).

## 1.6 Constant time step

All schemes up to now have been formulated for a general non-uniform mesh in time:  $t_0, t_1, \dots, t_{N_t}$ . Non-uniform meshes are highly relevant since one can use many points in regions where  $u$  varies rapidly, and save points in regions where  $u$  is slowly varying. This is the key idea of *adaptive* methods where the spacing of the mesh points are determined as the computations proceed.

However, a uniformly distributed set of mesh points is very common and sufficient for many applications. It therefore makes sense to present the finite difference schemes for a uniform point distribution  $t_n = n\Delta t$ , where  $\Delta t$  is the constant spacing between the mesh points, also referred to as the *time step*. The resulting formulas look simpler and are perhaps more well known.

**Summary of schemes for constant time step.**

$$u^{n+1} = (1 - a\Delta t)u^n \quad \text{Forward Euler} \quad (20)$$

$$u^{n+1} = \frac{1}{1 + a\Delta t}u^n \quad \text{Backward Euler} \quad (21)$$

$$u^{n+1} = \frac{1 - \frac{1}{2}a\Delta t}{1 + \frac{1}{2}a\Delta t}u^n \quad \text{Crank-Nicolson} \quad (22)$$

$$u^{n+1} = \frac{1 - (1 - \theta)a\Delta t}{1 + \theta a\Delta t}u^n \quad \text{The } \theta - \text{rule} \quad (23)$$

Not surprisingly, we present these three alternative schemes because they have different pros and cons, both for the simple ODE in question (which can easily be solved as accurately as desired), and for more advanced differential equation problems.

**Test the understanding.**

At this point it can be good training to apply the explained finite difference discretization techniques to a slightly different equation. Exercise 15 is therefore highly recommended to check that the key concepts are understood.

## 1.7 Mathematical derivation of finite difference formulas

The finite difference formulas for approximating the first derivative of a function have so far been somewhat justified through graphical illustrations in Figures 3, 4, and 5. The task is to approximate the derivative at a point of a curve using only two function values. By drawing a straight line through the points, we have some approximation to the tangent of the curve and use the slope of this line as an approximation to the derivative. The slope can be computed by inspecting the figures.

However, we can also derive the finite difference formulas by pure mathematics too. The key tool for this approach is Taylor series, or more precisely, approximation of functions by lower-order Taylor polynomials. Given a function  $f(x)$  with “enough derivatives”, a Taylor polynomial of degree  $m$  can be used to

approximate the value of the function  $f(x)$  if we know the values of  $f$  and its first  $m$  derivatives at some point  $x = a$ . The formula for the Taylor polynomial reads

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + \frac{1}{6}f'''(a)(x-a)^3 \dots + \frac{1}{m!} \frac{f^{(m)}}{x^m}(a)(x-a)^m. \quad (24)$$

For a function of time,  $f(t)$ , related to a mesh with spacing  $\Delta t$ , we often need the Taylor polynomial approximation at  $f(t_n \pm \Delta t)$  given  $f$  and its derivatives at  $t = t_n$ . Replacing  $x$  by  $t_n + \Delta t$  and  $a$  by  $t_n$  gives

$$f(t_n + \Delta t) \approx f(t_n) + f'(t_n)\Delta t + \frac{1}{2}f''(t_n)\Delta t^2 + \frac{1}{6}f'''(t_n)\Delta t^3 \dots + \frac{1}{m!} \frac{f^{(m)}}{x^m}(t_n)\Delta t^m. \quad (25)$$

**The forward difference.** We can use (25) to find an approximation for  $f'(t_n)$  simply by solving with respect to this quantity:

$$f'(t_n) \approx \frac{f(t_n + \Delta t) - f(t_n)}{\Delta t} - \frac{1}{2}f''(t_n)\Delta t - \frac{1}{6}f'''(t_n)\Delta t^2 \dots - \frac{1}{m!} \frac{f^{(m)}}{x^m}(t_n)\Delta t^{m-1}. \quad (26)$$

By letting  $m \rightarrow \infty$ , this formula is exact, but that is not so much of practical value. A more interesting observation is that all the power terms in  $\Delta t$  vanish as  $\Delta t \rightarrow 0$ , i.e., the formula

$$f'(t_n) \approx \frac{f(t_n + \Delta t) - f(t_n)}{\Delta t} \quad (27)$$

is exact in the limit  $\Delta t \rightarrow 0$ .

The interesting feature of (29) is that we have a measure of the error in the formula (27): the error is given by the extra terms on the right-hand side of (29). We assume that  $\Delta t$  is a small quantity. Then  $\Delta t^2 \ll \Delta t$ ,  $\Delta t^3 \ll \Delta t^2$ , and so on, which means that the first term is the dominating term. This first term reads  $-\frac{1}{2}f''(t_n)\Delta t$  and can be taken as a measure of the error in the Forward Euler formula.

**The backward difference.** To derive the backward difference, we use the Taylor polynomial approximation at  $f(t_n - \Delta t)$ :

$$f(t_n - \Delta t) \approx f(t_n) - f'(t_n)\Delta t + \frac{1}{2}f''(t_n)\Delta t^2 - \frac{1}{6}f'''(t_n)\Delta t^3 \dots + \frac{1}{m!} \frac{f^{(m)}}{x^m}(t_n)\Delta t^m. \quad (28)$$

Solving with respect to  $f'(t_n)$  gives

$$f'(t_n) \approx \frac{f(t_n) - f(t_n - \Delta t)}{\Delta t} + \frac{1}{2}f''(t_n)\Delta t - \frac{1}{6}f'''(t_n)\Delta t^2 \dots - \frac{1}{m!} \frac{f^{(m)}(t_n)}{x^m} \Delta t^{m-1}. \quad (29)$$

The term  $\frac{1}{2}f''(t_n)\Delta t$  can be taken as a simple measure of the approximation error since it will dominate over the other terms as  $\Delta t \rightarrow 0$ .

**The centered difference.** The centered difference approximates the derivative at  $t_n + \frac{1}{2}\Delta t$ . Let us write up the Taylor polynomial approximations to  $f(t_n)$  and  $f(t_{n+1})$  around  $t_n + \frac{1}{2}\Delta t$ :

$$f(t_n) \approx f(t_n + \frac{1}{2}\Delta t) - f'(t_n + \frac{1}{2}\Delta t)\frac{1}{2}\Delta t + f''(t_n + \frac{1}{2}\Delta t)(\frac{1}{2}\Delta t)^2 - f'''(t_n + \frac{1}{2}\Delta t)(\frac{1}{2}\Delta t)^3 + \dots \quad (30)$$

$$f(t_{n+1}) \approx f(t_n + \frac{1}{2}\Delta t) + f'(t_n + \frac{1}{2}\Delta t)\frac{1}{2}\Delta t + f''(t_n + \frac{1}{2}\Delta t)(\frac{1}{2}\Delta t)^2 + f'''(t_n + \frac{1}{2}\Delta t)(\frac{1}{2}\Delta t)^3 + \dots \quad (31)$$

Subtracting the first from the second gives

$$f(t_{n+1}) - f(t_n) = f'(t_n + \frac{1}{2}\Delta t)\Delta t + 2f'''(t_n + \frac{1}{2}\Delta t)(\frac{1}{2}\Delta t)^3 + \dots \quad (32)$$

Solving with respect to  $f'(t_n + \frac{1}{2}\Delta t)$  results in

$$f'(t_n + \frac{1}{2}\Delta t) \approx \frac{f(t_{n+1}) - f(t_n)}{\Delta t} - \frac{1}{4}f'''(t_n + \frac{1}{2}\Delta t)\Delta t^2 + c \dots \quad (33)$$

This time the error measure goes like  $\frac{1}{4}f''' \Delta t^2$ , i.e., it is proportional to  $\Delta t^2$  and not only  $\Delta t$ , which means that the error goes faster to zero as  $\Delta t$  is reduced. This means that the centered difference formula

$$f'(t_n + \frac{1}{2}\Delta t) \approx \frac{f(t_{n+1}) - f(t_n)}{\Delta t} \quad (34)$$

is more accurate than the forward and backward differences for small  $\Delta t$ .

## 1.8 Compact operator notation for finite differences

Finite difference formulas can be tedious to write and read, especially for differential equations with many terms and many derivatives. To save space and help the reader of the scheme to quickly see the nature of the difference approximations, we introduce a compact notation. A forward difference approximation is denoted by the  $D_t^+$  operator:

$$[D_t^+ u]^n = \frac{u^{n+1} - u^n}{\Delta t} \approx \frac{d}{dt}u(t_n). \quad (35)$$



The notation consists of an operator that approximates differentiation with respect to an independent variable, here  $t$ . The operator is built of the symbol  $D$ , with the variable as subscript and a superscript denoting the type of difference. The superscript  $+$  indicates a forward difference. We place square brackets around the operator and the function it operates on and specify the mesh point, where the operator is acting, by a superscript.

The corresponding operator notation for a centered difference and a backward difference reads

$$[D_t u]^n = \frac{u^{n+\frac{1}{2}} - u^{n-\frac{1}{2}}}{\Delta t} \approx \frac{d}{dt} u(t_n), \quad (36)$$

and

$$[D_t^- u]^n = \frac{u^n - u^{n-1}}{\Delta t} \approx \frac{d}{dt} u(t_n). \quad (37)$$

Note that the superscript  $-$  denotes the backward difference, while no superscript implies a central difference.

An averaging operator is also convenient to have:

$$[\bar{u}^t]^n = \frac{1}{2}(u^{n-\frac{1}{2}} + u^{n+\frac{1}{2}}) \approx u(t_n) \quad (38)$$

The superscript  $t$  indicates that the average is taken along the time coordinate. The common average  $(u^n + u^{n+1})/2$  can now be expressed as  $[\bar{u}^t]^{n+\frac{1}{2}}$ . (When also spatial coordinates enter the problem, we need the explicit specification of the coordinate after the bar.)

The Backward Euler finite difference approximation to  $u' = -au$  can be written as follows utilizing the compact notation:

$$[D_t^- u]^n = -au^n.$$

In difference equations we often place the square brackets around the whole equation, to indicate at which mesh point the equation applies, since each term is supposed to be approximated at the same point:

$$[D_t^- u = -au]^n. \quad (39)$$

The Forward Euler scheme takes the form

$$[D_t^+ u = -au]^n, \quad (40)$$

while the Crank-Nicolson scheme is written as

$$[D_t u = -a\bar{u}^t]^{n+\frac{1}{2}}. \quad (41)$$

**Question.**

Apply (36) and (38) and write out the expressions to see that (41) is indeed the Crank-Nicolson scheme.

The  $\theta$ -rule can be specified by

$$[\bar{D}_t u = -a\bar{u}^{t,\theta}]^{n+\theta}, \quad (42)$$

if we define a new time difference

$$[\bar{D}_t u]^{n+\theta} = \frac{u^{n+1} - u^n}{t^{n+1} - t^n}, \quad (43)$$

and a *weighted averaging operator*

$$[\bar{u}^{t,\theta}]^{n+\theta} = (1 - \theta)u^n + \theta u^{n+1} \approx u(t_{n+\theta}), \quad (44)$$

where  $\theta \in [0, 1]$ . Note that for  $\theta = \frac{1}{2}$  we recover the standard centered difference and the standard arithmetic mean. The idea in (42) is to sample the equation at  $t_{n+\theta}$ , use a skew difference at that point  $[\bar{D}_t u]^{n+\theta}$ , and a skew mean value. An alternative notation is

$$[D_t u]^{n+\frac{1}{2}} = \theta[-au]^{n+1} + (1 - \theta)[-au]^n.$$

Looking at the various examples above and comparing them with the underlying differential equations, we see immediately which difference approximations that have been used and at which point they apply. Therefore, the compact notation effectively communicates the reasoning behind turning a differential equation into a difference equation.

## 2 Implementation

### Goal.

We want make a computer program for solving

$$u'(t) = -au(t), \quad t \in (0, T], \quad u(0) = I,$$

by finite difference methods. The program should also display the numerical solution as a curve on the screen, preferably together with the exact solution.

All programs referred to in this section are found in the [src/decay](#) directory (we use the classical Unix term *directory* for what many others nowadays call *folder*).

**Mathematical problem.** We want to explore the Forward Euler scheme, the Backward Euler, and the Crank-Nicolson schemes applied to our model problem. From an implementational point of view, it is advantageous to implement the  $\theta$ -rule

$$u^{n+1} = \frac{1 - (1 - \theta)a\Delta t}{1 + \theta a\Delta t} u^n,$$

since it can generate the three other schemes by various choices of  $\theta$ :  $\theta = 0$  for Forward Euler,  $\theta = 1$  for Backward Euler, and  $\theta = 1/2$  for Crank-Nicolson. Given  $a$ ,  $u^0 = I$ ,  $T$ , and  $\Delta t$ , our task is to use the  $\theta$ -rule to compute  $u^1, u^2, \dots, u^{N_t}$ , where  $t_{N_t} = N_t \Delta t$ , and  $N_t$  the closest integer to  $T/\Delta t$ .

**Computer Language: Python.** Any programming language can be used to generate the  $u^{n+1}$  values from the formula above. However, in this document we shall mainly make use of Python of several reasons:

- Python has a very clean, readable syntax (often known as "executable pseudo-code").
- Python code is very similar to MATLAB code (and MATLAB has a particularly widespread use for scientific computing).
- Python is a full-fledged, very powerful programming language.
- Python is similar to, but much simpler to work with and results in more reliable code than C++.
- Python has a rich set of modules for scientific computing, and its popularity in scientific computing is rapidly growing.
- Python was made for being combined with compiled languages (C, C++, Fortran) to reuse existing numerical software and to reach high computational performance of new implementations.
- Python has extensive support for administrative task needed when doing large-scale computational investigations.
- Python has extensive support for graphics (visualization, user interfaces, web applications).
- FEniCS, a very powerful tool for solving PDEs by the finite element method, is most human-efficient to operate from Python.

Learning Python is easy. Many newcomers to the language will probably learn enough from the forthcoming examples to perform their own computer experiments. The examples start with simple Python code and gradually make use of more powerful constructs as we proceed. As long as it is not inconvenient for the problem at hand, our Python code is made as close as possible to MATLAB code for easy transition between the two languages.

Readers who feel the Python examples are too hard to follow will probably benefit from reading a tutorial, e.g.,

- [The Official Python Tutorial](#)
- [Python Tutorial on tutorialspoint.com](#)
- [Interactive Python tutorial site](#)
- [A Beginner's Python Tutorial](#)

The author also has a comprehensive book [4] that teaches scientific programming with Python from the ground up.

## 2.1 Making a solver function

We choose to have an array  $u$  for storing the  $u^n$  values,  $n = 0, 1, \dots, N_t$ . The algorithmic steps are

1. initialize  $u^0$
2. for  $t = t_n$ ,  $n = 1, 2, \dots, N_t$ : compute  $u_n$  using the  $\theta$ -rule formula

**Function for computing the numerical solution.** The following Python function takes the input data of the problem  $(I, a, T, \Delta t, \theta)$  as arguments and returns two arrays with the solution  $u^0, \dots, u^{N_t}$  and the mesh points  $t_0, \dots, t_{N_t}$ , respectively:

```
from numpy import *

def solver(I, a, T, dt, theta):
    """Solve u'=-a*u, u(0)=I, for t in (0,T] with steps of dt."""
    Nt = int(T/dt)          # no of time intervals
    T = Nt*dt              # adjust T to fit time step dt
    u = zeros(Nt+1)         # array of u[n] values
    t = linspace(0, T, Nt+1) # time mesh

    u[0] = I                # assign initial condition
    for n in range(0, Nt):  # n=0,1,...,Nt-1
        u[n+1] = (1 - (1-theta)*a*dt)/(1 + theta*dt*a)*u[n]
    return u, t
```

The `numpy` library contains a lot of functions for array computing. Most of the function names are similar to what is found in the alternative scientific computing language MATLAB. Here we make use of

- `zeros(Nt+1)` for creating an array of a size  $Nt+1$  and initializing the elements to zero
- `linspace(0, T, Nt+1)` for creating an array with  $Nt+1$  coordinates uniformly distributed between 0 and T

The `for` loop deserves a comment, especially for newcomers to Python. The construction `range(0, Nt, s)` generates all integers from 0 to `Nt` in steps of `s`, *but not including* `Nt`. Omitting `s` means `s=1`. For example, `range(0, 6, 3)` gives 0 and 3, while `range(0, Nt)` generates 0, 1, ..., `Nt-1`. Our loop implies the following assignments to `u[n+1]`: `u[1]`, `u[2]`, ..., `u[Nt]`, which is what we want since `u` has length `Nt+1`. The first index in Python arrays or lists is *always* 0 and the last is then `len(u)-1`. The length of an array `u` is obtained by `len(u)` or `u.size`.

To compute with the `solver` function, we need to *call* it. Here is a sample call:

```
u, t = solver(I=1, a=2, T=8, dt=0.8, theta=1)
```

**Integer division.** The shown implementation of the `solver` may face problems and wrong results if `T`, `a`, `dt`, and `theta` are given as integers, see Exercises 2 and 3. The problem is related to *integer division* in Python (as well as in Fortran, C, C++, and many other computer languages):  $1/2$  becomes 0, while  $1.0/2$ ,  $1/2.0$ , or  $1.0/2.0$  all become 0.5. It is enough that at least the nominator or the denominator is a real number (i.e., a `float` object) to ensure correct mathematical division. Inserting a conversion `dt = float(dt)` guarantees that `dt` is `float` and avoids problems in Exercise 3.

Another problem with computing  $N_t = T/\Delta t$  is that we should round  $N_t$  to the nearest integer. With `Nt = int(T/dt)` the `int` operation picks the largest integer smaller than  $T/\Delta t$ . Correct mathematical rounding as known from school is obtained by

```
Nt = int(round(T/dt))
```

The complete version of our improved, safer `solver` function then becomes

```
from numpy import *

def solver(I, a, T, dt, theta):
    """Solve u'=-a*u, u(0)=I, for t in (0,T] with steps of dt."""
    dt = float(dt)          # avoid integer division
    Nt = int(round(T/dt))    # no of time intervals
    T = Nt*dt              # adjust T to fit time step dt
    u = zeros(Nt+1)        # array of u[n] values
    t = linspace(0, T, Nt+1) # time mesh

    u[0] = I               # assign initial condition
    for n in range(0, Nt): # n=0,1,...,Nt-1
        u[n+1] = (1 - (1-theta)*a*dt)/(1 + theta*dt*a)*u[n]
    return u, t
```

**Doc strings.** Right below the header line in the `solver` function there is a Python string enclosed in triple double quotes `"""`. The purpose of this string

object is to document what the function does and what the arguments are. In this case the necessary documentation do not span more than one line, but with triple double quoted strings the text may span several lines:

```
def solver(I, a, T, dt, theta):
    """
    Solve

        u'(t) = -a*u(t),

    with initial condition u(0)=I, for t in the time interval
    (0,T]. The time interval is divided into time steps of
    length dt.

    theta=1 corresponds to the Backward Euler scheme, theta=0
    to the Forward Euler scheme, and theta=0.5 to the Crank-
    Nicolson method.
    """
    ...
```

Such documentation strings appearing right after the header of a function are called *doc strings*. There are tools that can automatically produce nicely formatted documentation by extracting the definition of functions and the contents of doc strings.

It is strongly recommended to equip any function whose purpose is not obvious with a doc string. Nevertheless, the forthcoming text deviates from this rule if the function is explained in the text.

**Formatting of numbers.** Having computed the discrete solution  $u$ , it is natural to look at the numbers:

```
# Write out a table of t and u values:
for i in range(len(t)):
    print t[i], u[i]
```

This compact `print` statement gives unfortunately quite ugly output because the `t` and `u` values are not aligned in nicely formatted columns. To fix this problem, we recommend to use the *printf format*, supported most programming languages inherited from C. Another choice is Python's recent *format string syntax*.

Writing `t[i]` and `u[i]` in two nicely formatted columns is done like this with the `printf` format:

```
print 't=%6.3f u=%g' % (t[i], u[i])
```

The percentage signs signify "slots" in the text where the variables listed at the end of the statement are inserted. For each "slot" one must specify a format for how the variable is going to appear in the string: `s` for pure text, `d` for an integer, `g` for a real number written as compactly as possible, `9.3E` for scientific notation with three decimals in a field of width 9 characters (e.g., `-1.351E-2`), or `.2f` for standard decimal notation with two decimals formatted with minimum width.

The printf syntax provides a quick way of formatting tabular output of numbers with full control of the layout.

The alternative *format string syntax* looks like

```
print 't={t:6.3f} u={u:g}'.format(t=t[i], u=u[i])
```

As seen, this format allows logical names in the "slots" where `t[i]` and `u[i]` are to be inserted. The "slots" are surrounded by curly braces, and the logical name is followed by a colon and then the printf-like specification of how to format real numbers, integers, or strings.

**Running the program.** The function and main program shown above must be placed in a file, say with name `decay_v1.py` (v1 for 1st version of this program). Make sure you write the code with a suitable text editor (Gedit, Emacs, Vim, Notepad++, or similar). The program is run by executing the file this way:

---

```
Terminal> python decay_v1.py
```

---

The text `Terminal>` just indicates a prompt in a Unix/Linux or DOS terminal window. After this prompt, which will look different in your terminal window, depending on the terminal application and how it is set up, commands like `python decay_v1.py` can be issued. These commands are interpreted by the operating system.

We strongly recommend to run Python programs within the IPython shell. First start IPython by typing `ipython` in the terminal window. Inside the IPython shell, our program `decay_v1.py` is run by the command `run decay_v1.py`:

---

```
Terminal> ipython

In [1]: run decay_v1.py
t= 0.000 u=1
t= 0.800 u=0.384615
t= 1.600 u=0.147929
t= 2.400 u=0.0568958
t= 3.200 u=0.021883
t= 4.000 u=0.00841653
t= 4.800 u=0.00323713
t= 5.600 u=0.00124505
t= 6.400 u=0.000478865
t= 7.200 u=0.000184179
t= 8.000 u=7.0838e-05
```

```
In [2]:
```

---

The advantage of running programs in IPython are many: previous commands are easily recalled with the up arrow, `%pdb` turns on debugging so that variables

can be examined if the program aborts due to an exception, output of commands are stored in variables, programs and statements can be profiled, any operating system command can be executed, modules can be loaded automatically and other customizations can be performed when starting IPython – to mention a few of the most useful features.

Although running programs in IPython is strongly recommended, most execution examples in the forthcoming text use the standard Python shell with prompt `>>>` and run programs through a typesetting like

---

```
Terminal> python programname
```

---

The reason is that such typesetting makes the text more compact in the vertical direction than showing sessions with IPython syntax.

**Plotting the solution.** Having the `t` and `u` arrays, the approximate solution `u` is visualized by the intuitive command `plot(t, u)`:

```
from matplotlib.pyplot import *
plot(t, u)
show()
```

It will be illustrative to also plot  $u_e(t)$  for comparison. We first need to make a function for computing the analytical solution  $u_e(t) = Ie^{-at}$  of the model problem:

```
def exact_solution(t, I, a):
    return I*exp(-a*t)
```

It is tempting to just do

```
u_e = exact_solution(t, I, a)
plot(t, u, t, u_e)
```

However, this is not exactly what we want: the `plot` function draws straight lines between the discrete points `(t[n], u_e[n])` while  $u_e(t)$  varies as an exponential function between the mesh points. The technique for showing the “exact” variation of  $u_e(t)$  between the mesh points is to introduce a very fine mesh for  $u_e(t)$ :

```
t_e = linspace(0, T, 1001)    # fine mesh
u_e = exact_solution(t_e, I, a)
```

We can also plot the curves with different colors and styles, e.g.,



```

plot(t_e, u_e, 'b-',      # blue line for u_e
     t,  u,   'r--o')    # red dashes w/circles

```

With more than one curve in the plot we need to associate each curve with a legend. We also want appropriate names on the axis, a title, and a file containing the plot as an image for inclusion in reports. The Matplotlib package (`matplotlib.pyplot`) contains functions for this purpose. The names of the functions are similar to the plotting functions known from MATLAB. A complete function for creating the comparison plot becomes

```

from matplotlib.pyplot import *

def plot_numerical_and_exact(theta, I, a, T, dt):
    """Compare the numerical and exact solution in a plot."""
    u, t = solver(I=I, a=a, T=T, dt=dt, theta=theta)

    t_e = linspace(0, T, 1001)      # fine mesh for u_e
    u_e = exact_solution(t_e, I, a)

    plot(t,  u,   'r--o',            # red dashes w/circles
         t_e, u_e, 'b-',            # blue line for exact sol.
         legend(['numerical', 'exact'])
    xlabel('t')
    ylabel('u')
    title('theta=%g, dt=%g' % (theta, dt))
    savefig('plot_%s_%g.png' % (theta, dt))

plot_numerical_and_exact(I=1, a=2, T=8, dt=0.8, theta=1)
show()

```

Note that `savefig` here creates a PNG file whose name reflects the values of  $\theta$  and  $\Delta t$  so that we can easily distinguish files from different runs with  $\theta$  and  $\Delta t$ .

The complete code is found in the file `decay_v2.py`. The resulting plot is shown in Figure 6. As seen, there is quite some discrepancy between the exact and the numerical solution. Fortunately, the numerical solution approaches the exact one as  $\Delta t$  is reduced.

## 2.2 Verifying the implementation

It is easy to make mistakes while deriving and implementing numerical algorithms, so we should never believe in the solution before it has been thoroughly verified. The most obvious idea to verify the computations is to compare the numerical solution with the exact solution, when that exists, but there will always be a discrepancy between these two solutions because of the numerical approximations. The challenging question is whether we have the mathematically correct discrepancy or if we have another, maybe small, discrepancy due to both an approximation error and an error in the implementation. When looking at Figure 6, it is impossible to judge whether the program is correct or not.

The purpose of *verifying* a program is to bring evidence for the property that there are no errors in the implementation. To avoid mixing unavoidable approximation errors and undesired implementation errors, we should try to

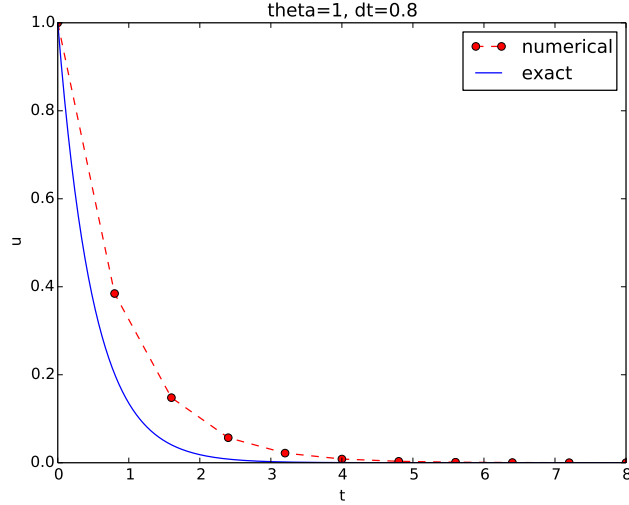


Figure 6: Comparison of numerical and exact solution.

make tests where we have some exact computation of the discrete solution or at least parts of it. Examples will show how this can be done.

**Running a few algorithmic steps by hand.** The simplest approach to produce a correct non-trivial reference solution for the discrete solution  $u$  of finite difference equations is to compute a few steps of the algorithm by hand. Then we can compare the hand calculations with numbers produced by the program.

A straightforward approach is to use a calculator and compute  $u^1$ ,  $u^2$ , and  $u^3$ . With  $I = 0.1$ ,  $\theta = 0.8$ , and  $\Delta t = 0.8$  we get

$$A \equiv \frac{1 - (1 - \theta)a\Delta t}{1 + \theta a\Delta t} = 0.298245614035$$

$$u^1 = AI = 0.0298245614035,$$

$$u^2 = Au^1 = 0.00889504462912,$$

$$u^3 = Au^2 = 0.00265290804728$$

Comparison of these manual calculations with the result of the `solver` function is carried out in the function

```
def test_solver_three_steps():
    """Compare three steps with known manual computations."""
    theta = 0.8; a = 2; I = 0.1; dt = 0.8
    u_by_hand = array([I,
```

```

0.0298245614035,
0.00889504462912,
0.00265290804728])

Nt = 3 # number of time steps
u, t = solver(I=I, a=a, T=Nt*dt, dt=dt, theta=theta)

tol = 1E-15 # tolerance for comparing floats
diff = abs(u - u_by_hand).max()
success = diff <= tol
assert success

```

The `test_solver_three_steps` function follows widely used conventions for *unit testing*. By following such conventions we can at a later stage easily execute a big test suite for our software. The conventions are three-fold:

- The test function starts with `test_` and takes no arguments.
- The test ends up in a boolean expression that is `True` if the test passed and `False` if it failed.
- The function runs `assert` on the boolean expression, resulting in program abortion (due to an `AssertionError` exception) if the test failed.

The main program can routinely run the verification test prior to solving the real problem:

```

test_solver_three_steps()
plot_numerical_and_exact(I=1, a=2, T=8, dt=0.8, theta=1)
show()

```

(Rather than calling `test_*` functions explicitly, one will normally ask a testing framework like nose or pytest to find and run such functions.) The complete program including the verification above is found in the file `decay_v3.py`.

## 2.3 Computing the numerical error as a mesh function

Now that we have some evidence for a correct implementation, we are in a position to compare the computed  $u^n$  values in the `u` array with the exact  $u$  values at the mesh points, in order to study the error in the numerical solution.

A natural way to compare the exact and discrete solutions is to calculate their difference as a mesh function:

$$e^n = u_e(t_n) - u^n, \quad n = 0, 1, \dots, N_t. \quad (45)$$

We may view  $u_e^n = u_e(t_n)$  as the representation of  $u_e(t)$  as a mesh function rather than a continuous function defined for all  $t \in [0, T]$  ( $u_e^n$  is often called the *representative* of  $u_e$  on the mesh). Then,  $e^n = u_e^n - u^n$  is clearly the difference of two mesh functions. This interpretation of  $e^n$  is natural when programming.

The error mesh function  $e^n$  can be computed by

```

u, t = solver(I, a, T, dt, theta) # Numerical sol.
u_e = exact_solution(t, I, a)     # Representative of exact sol.
e = u_e - u

```

Note that the mesh functions  $u$  and  $u_e$  are represented by arrays and associated with the points in the array  $t$ .

#### Array arithmetics.

The last statements

```

u_e = exact_solution(t, I, a)
e = u_e - u

```

are primary examples of array arithmetics:  $t$  is an array of mesh points that we pass to `exact_solution`. This function evaluates `-a*t`, which is a scalar times an array, meaning that the scalar is multiplied with each array element. The result is an array, let us call it `tmp1`. Then `exp(tmp1)` means applying the exponential function to each element in `tmp1`, resulting an array, say `tmp2`. Finally, `I*tmp2` is computed (scalar times array) and  $u_e$  refers to this array returned from `exact_solution`. The expression `u_e - u` is the difference between two arrays, resulting in a new array referred to by `e`.

## 2.4 Computing the norm of the numerical error

Instead of working with the error  $e^n$  on the entire mesh, we often want one number expressing the size of the error. This is obtained by taking the norm of the error function.

Let us first define norms of a function  $f(t)$  defined for all  $t \in [0, T]$ . Three common norms are

$$\|f\|_{L^2} = \left( \int_0^T f(t)^2 dt \right)^{1/2}, \quad (46)$$

$$\|f\|_{L^1} = \int_0^T |f(t)| dt, \quad (47)$$

$$\|f\|_{L^\infty} = \max_{t \in [0, T]} |f(t)|. \quad (48)$$

The  $L^2$  norm (46) ("L-two norm") has nice mathematical properties and is the most popular norm. It is a generalization of the well-known Euclidian norm of vectors to functions. The  $L^\infty$  is also called the max norm or the supremum norm. In fact, there is a whole family of norms,

$$\|f\|_{L^p} = \left( \int_0^T f(t)^p dt \right)^{1/p}, \quad (49)$$

with  $p$  real. In particular,  $p = 1$  corresponds to the  $L^1$  norm above while  $p = \infty$  is the  $L^\infty$  norm.

Numerical computations involving mesh functions need corresponding norms. Given a set of function values,  $f^n$ , and some associated mesh points,  $t_n$ , a numerical integration rule can be used to calculate the  $L^2$  and  $L^1$  norms defined above. Imagining that the mesh function is extended to vary linearly between the mesh points, the Trapezoidal rule is in fact an exact integration rule. A possible modification of the  $L^2$  norm for a mesh function  $f^n$  on a uniform mesh with spacing  $\Delta t$  is therefore the well-known Trapezoidal integration formula

$$\|f^n\| = \left( \Delta t \left( \frac{1}{2}(f^0)^2 + \frac{1}{2}(f^{N_t})^2 + \sum_{n=1}^{N_t-1} (f^n)^2 \right) \right)^{1/2}$$

A common approximation of this expression, motivated by the convenience of having a simpler formula, is

$$\|f^n\|_{\ell^2} = \left( \Delta t \sum_{n=0}^{N_t} (f^n)^2 \right)^{1/2}.$$

This is called the discrete  $L^2$  norm and denoted by  $\ell^2$ . The error in  $\|f\|_{\ell^2}^2$  compared with the Trapezoidal integration formula is  $\Delta t((f^0)^2 + (f^{N_t})^2)/2$ , which means perturbed weights at the end points of the mesh function, and the error goes to zero as  $\Delta t \rightarrow 0$ . As long as we are consistent and stick to one kind of integration rule for the norm of a mesh function, the details and accuracy of this rule is not of concern.

The three discrete norms for a mesh function  $f^n$ , corresponding to the  $L^2$ ,  $L^1$ , and  $L^\infty$  norms of  $f(t)$  defined above, are defined by

$$\|f^n\|_{\ell^2} = \left( \Delta t \sum_{n=0}^{N_t} (f^n)^2 \right)^{1/2}, \quad (50)$$

$$\|f^n\|_{\ell^1} = \Delta t \sum_{n=0}^{N_t} |f^n| \quad (51)$$

$$\|f^n\|_{\ell^\infty} = \max_{0 \leq n \leq N_t} |f^n|. \quad (52)$$

Note that the  $L^2$ ,  $L^1$ ,  $\ell^2$ , and  $\ell^1$  norms depend on the length of the interval of interest (think of  $f = 1$ , then the norms are proportional to  $\sqrt{T}$  or  $T$ ). In some applications it is convenient to think of a mesh function as just a vector of function values and neglect the information of the mesh points. Then we can replace  $\Delta t$  by  $T/N_t$  and drop  $T$ . Moreover, it is convenient to divide by the

total length of the vector,  $N_t + 1$ , instead of  $N_t$ . This reasoning gives rise to the *vector norms* for a vector  $f = (f_0, \dots, f_N)$ :

$$\|f\|_2 = \left( \frac{1}{N+1} \sum_{n=0}^N (f_n)^2 \right)^{1/2}, \quad (53)$$

$$\|f\|_1 = \frac{1}{N+1} \sum_{n=0}^N |f_n| \quad (54)$$

$$\|f\|_{\ell^\infty} = \max_{0 \leq n \leq N} |f_n|. \quad (55)$$

Here we have used the common vector component notation with subscripts ( $f_n$ ) and  $N$  as length. We will mostly work with mesh functions and use the discrete  $\ell^2$  norm (50) or the max norm  $\ell^\infty$  (52), but the corresponding vector norms (53)-(55) are also much used in numerical computations, so it is important to know the different norms and the relations between them.

A single number that expresses the size of the numerical error will be taken as  $\|e^n\|_{\ell^2}$  and called  $E$ :

$$E = \sqrt{\Delta t \sum_{n=0}^{N_t} (e^n)^2} \quad (56)$$

The corresponding Python code, using array arithmetics, reads

```
E = sqrt(dt*sum(e**2))
```

The `sum` function comes from `numpy` and computes the sum of the elements of an array. Also the `sqrt` function is from `numpy` and computes the square root of each element in the array argument.

**Scalar computing.** Instead of doing array computing `sqrt(dt*sum(e**2))` we can compute with one element at a time:

```
m = len(u)      # length of u array (alt: u.size)
u_e = zeros(m)
t = 0
for i in range(m):
    u_e[i] = exact_solution(t, a, I)
    t = t + dt
e = zeros(m)
for i in range(m):
    e[i] = u_e[i] - u[i]
s = 0 # summation variable
for i in range(m):
    s = s + e[i]**2
error = sqrt(dt*s)
```

Such element-wise computing, often called *scalar* computing, takes more code, is less readable, and runs much slower than what we can achieve with array computing.

## 2.5 Experiments with computing and plotting

Let us wrap up the computation of the error measure and all the plotting statements for comparing the exact and numerical solution in a new function `explore`. This function can be called for various  $\theta$  and  $\Delta t$  values to see how the error varies with the method and the mesh resolution:

```
def explore(I, a, T, dt, theta=0.5, makeplot=True):
    """
    Run a case with the solver, compute error measure,
    and plot the numerical and exact solutions (if makeplot=True).
    """
    u, t = solver(I, a, T, dt, theta)    # Numerical solution
    u_e = exact_solution(t, I, a)
    e = u_e - u
    E = sqrt(dt*sum(e**2))
    if makeplot:
        figure()                          # create new plot
        t_e = linspace(0, T, 1001)       # fine mesh for u_e
        u_e = exact_solution(t_e, I, a)
        plot(t, u, 'r--o')                # red dashes w/circles
        plot(t_e, u_e, 'b-')              # blue line for exact sol.
        legend(['numerical', 'exact'])
        xlabel('t')
        ylabel('u')
        title('theta=%g, dt=%g' % (theta, dt))
        theta2name = {0: 'FE', 1: 'BE', 0.5: 'CN'}
        savefig('%s_%g.png' % (theta2name[theta], dt))
        savefig('%s_%g.pdf' % (theta2name[theta], dt))
        show()
    return E
```

The `figure()` call is key: without it, a new `plot` command will draw the new pair of curves in the same plot window, while we want the different pairs to appear in separate windows and files. Calling `figure()` ensures this.

Filenames with the method name (FE, BE, or CN) rather than the  $\theta$  value embedded in the name, can easily be created with the aid of a little Python dictionary for mapping  $\theta$  to method acronyms:

```
theta2name = {0: 'FE', 1: 'BE', 0.5: 'CN'}
savefig('%s_%g.png' % (theta2name[theta], dt))
```

The `explore` function stores the plot in two different image file formats: PNG and PDF. The PNG format is aimed at being included in HTML files and the PDF format in  $\text{\LaTeX}$  documents (more precisely, in PDF $\text{\LaTeX}$  documents). Frequently used viewers for these image files on Unix systems are `gv` (comes with Ghostscript) for the PDF format and `display` (from the ImageMagick suite for PNG files:

---

```
Terminal> gv BE_0.5.pdf
Terminal> display BE_0.5.png
```

---

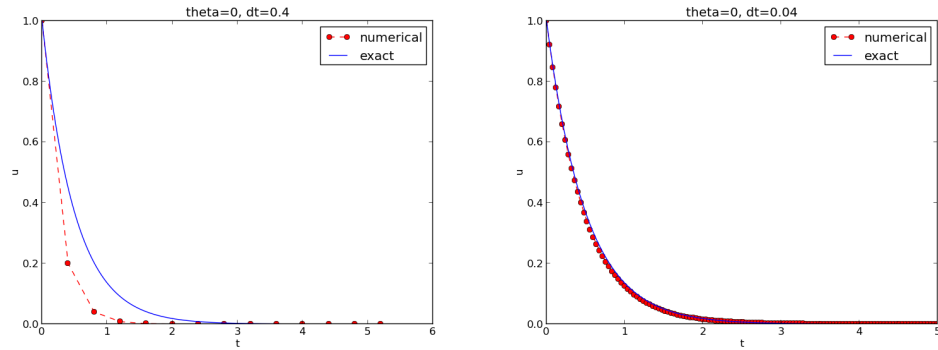


Figure 7: The Forward Euler scheme for two values of the time step.

A main program may run a loop over the three methods ( $\theta$  values) and call `explore` to compute errors and make plots:

```
def main(I, a, T, dt_values, theta_values=(0, 0.5, 1)):
    for theta in theta_values:
        for dt in dt_values:
            E = explore(I, a, T, dt, theta, makeplot=True)
            print '%3.1f %6.2f: %12.3E' % (theta, dt, E)
```

The complete code containing the functions above resides in the file `decay_plot_mpl.py`. Running this program results in

---

```
Terminal> python decay_plot_mpl.py
0.0  0.40:    2.105E-01
0.0  0.04:    1.449E-02
0.5  0.40:    3.362E-02
0.5  0.04:    1.887E-04
1.0  0.40:    1.030E-01
1.0  0.04:    1.382E-02
```

---

We observe that reducing  $\Delta t$  by a factor of 10 increases the accuracy for all three methods ( $\theta$  values). We also see that the combination of  $\theta = 0.5$  and a small time step  $\Delta t = 0.04$  gives a much more accurate solution, and that  $\theta = 0$  and  $\theta = 1$  with  $\Delta t = 0.4$  result in the least accurate solutions.

Figure 7 demonstrates that the numerical solution for  $\Delta t = 0.4$  clearly lies below the exact curve, but that the accuracy improves considerably by reducing the time step by a factor of 10.

**Combining plot files.** Mounting two PNG files, as done in the figure, is easily done by the `montage` program from the ImageMagick suite:



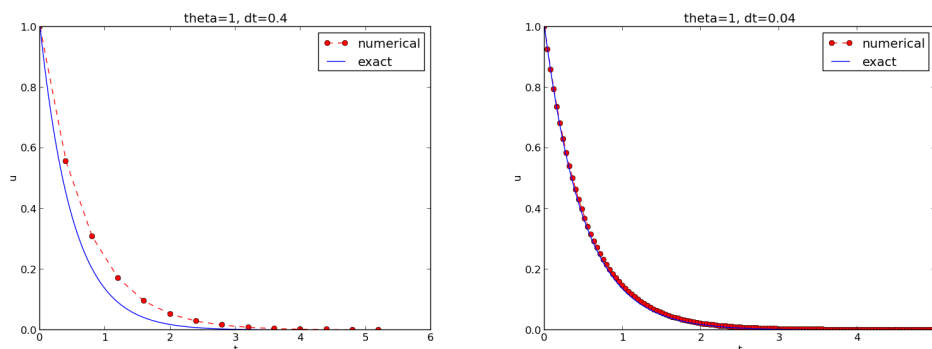


Figure 8: The Backward Euler scheme for two values of the time step.

---

```
Terminal> montage -background white -geometry 100% -tile 2x1 \
    FE_0.4.png FE_0.04.png FE1.png
Terminal> convert -trim FE1.png FE1.png
```

---

The `-geometry` argument is used to specify the size of the image, and here we preserve the individual sizes of the images. The `-tile HxV` option specifies `H` images in the horizontal direction and `V` images in the vertical direction. A series of image files to be combined are then listed, with the name of the resulting combined image, here `FE1.png` at the end. The `convert -trim` command removes surrounding white areas in the figure (an operation usually known as *cropping* in image manipulation programs).

For  $\text{\LaTeX}$  reports it is not recommended to use `montage` and PNG files as the result has too low resolution. Instead, plots should be made in the PDF format and combined using the `pdftk`, `pdfnup`, and `pdfcrop` tools (on Linux/Unix):

---

```
Terminal> pdftk FE_0.4.png FE_0.04.png output tmp.pdf
Terminal> pdfnup --nup 2x1 --outfile tmp.pdf tmp.pdf
Terminal> pdfcrop tmp.pdf FE1.png # output in FE1.png
```

---

Here, `pdftk` combines images into a multi-page PDF file, `pdfnup` combines the images in individual pages to a table of images (pages), and `pdfcrop` removes white margins in the resulting combined image file.

The behavior of the two other schemes is shown in Figures 8 and 9. Crank-Nicolson is obviously the most accurate scheme from this visual point of view.

**Plotting with SciTools.** The [SciTools package](#) provides a unified plotting interface, called Easyviz, to many different plotting packages, including Matplotlib, Gnuplot, Grace, MATLAB, VTK, OpenDX, and VisIt. The syntax is

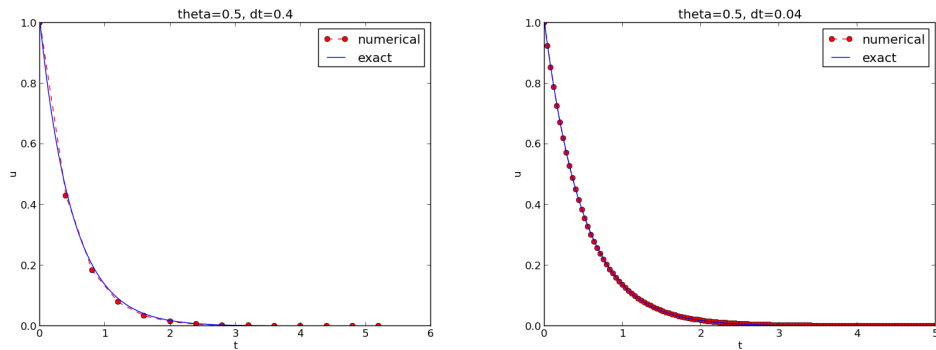


Figure 9: The Crank-Nicolson scheme for two values of the time step.

very similar to that of Matplotlib and MATLAB. In fact, the plotting commands shown above look the same in SciTool's Easyviz interface, apart from the import statement, which reads

```
from scitools.std import *
```

This statement performs a `from numpy import *` as well as an import of the most common pieces of the Easyviz (`scitools.easyviz`) package, along with some additional numerical functionality.

With Easyviz one can merge several plotting commands into a single one using keyword arguments:

```
plot(t, u, 'r--o',          # red dashes w/circles
     t_e, u_e, 'b-',        # blue line for exact sol.
     legend=['numerical', 'exact'],
     xlabel='t',
     ylabel='u',
     title='theta=%g, dt=%g' % (theta, dt),
     savefig='%s_%g.png' % (theta2name[theta], dt),
     show=True)
```

The `decay_plot_st.py` file contains such a demo.

By default, Easyviz employs Matplotlib for plotting, but [Gnuplot](#) and [Grace](#) are viable alternatives:

---

```
Terminal> python decay_plot_st.py --SCITOOLS_easyviz_backend gnuplot
Terminal> python decay_plot_st.py --SCITOOLS_easyviz_backend grace
```

---

The backend used for creating plots (and numerous other options) can be permanently set in SciTool's configuration file.

All the Gnuplot windows are launched without any need to kill one before the next one pops up (as is the case with Matplotlib) and one can press the key

'q' anywhere in a plot window to kill it. Another advantage of Gnuplot is the automatic choice of sensible and distinguishable line types in black-and-white PDF and PostScript files.

Regarding functionality for annotating plots with title, labels on the axis, legends, etc., we refer to the documentation of Matplotlib and SciTools for more detailed information on the syntax. The hope is that the programming syntax explained so far suffices for understanding the code and learning more from a combination of the forthcoming examples and other resources such as books and web pages.

### Test the understanding.

Exercise 16 asks you to implement a solver for a problem that is slightly different from the one above. You may use the `solver` and `explore` functions explained above as a starting point. Apply the new solver to Exercise 17.

## 2.6 Memory-saving implementation

The computer memory requirements of our implementations so far consists mainly of the `u` and `t` arrays, both of length  $N_t + 1$ , plus some other temporary arrays that Python needs for intermediate results if we do array arithmetics in our program (e.g., `I*exp(-a*t)` needs to store `a*t` before `-` can be applied to it and then `exp`). Regardless of how we implement simple ODE problems, storage requirements are very modest and put not restriction on how we choose our data structures and algorithms. Nevertheless, when the methods for ODEs used here are applied to three-dimensional partial differential equation (PDE) problems, memory storage requirements suddenly become a challenging issue.

The PDE counterpart to our model problem  $u' = -a$  is a diffusion equation  $u_t = a\nabla^2 u$  posed on a space-time domain. The discrete representation of this domain may in 3D be a spatial mesh of  $M^3$  points and a time mesh of  $N_t$  points. A typical desired value for  $M$  is 100 in many applications, or even 1000. Storing all the computed  $u$  values, like we have done in the programs so far, demands storage of some arrays of size  $M^3 N_t$ , giving a factor of  $M^3$  larger storage demands compared to our ODE programs. Each real number in the array for  $u$  requires 8 bytes (b) of storage. With  $M = 100$  and  $N_t = 1000$ , there is a storage demand of  $(10^3)^3 \cdot 1000 \cdot 8 = 8$  Gb for the solution array. Fortunately, we can usually get rid of the  $N_t$  factor, resulting in 8 Mb of storage. Below we explain how this is done, and the technique is almost always applied in implementations of PDE problems.

Let us critically evaluate how much we really need to store in the computer's memory in our implementation of the  $\theta$  method. To compute a new  $u^{n+1}$ , all we need is  $u^n$ . This implies that the previous  $u^{n-1}, u^{n-2}, \dots, u^0$  values do not need to be stored in an array, although this is convenient for plotting and data analysis

in the program. Instead of the `u` array we can work with two variables for real numbers, `u` and `u_1`, representing  $u^{n+1}$  and  $u^n$  in the algorithm, respectively. At each time level, we update `u` from `u_1` and then set `u_1 = u` so that the computed  $u^{n+1}$  value becomes the "previous" value  $u^n$  at the next time level. The downside is that we cannot plot the solution after the simulation is done since only the last two numbers are available. The remedy is to store computed values in a file and use the file for visualizing the solution later.

We have implemented this memory saving idea in the file `decay_memsave.py`, which is a slight modification of `decay_plot_mpl.py` program.

The following function demonstrates how we work with the two most recent values of the unknown:

```
def solver_memsave(I, a, T, dt, theta, filename='sol.dat'):
    """
    Solve u'=-a*u, u(0)=I, for t in (0,T] with steps of dt.
    Minimum use of memory. The solution is stored in a file
    (with name filename) for later plotting.
    """
    dt = float(dt)          # avoid integer division
    Nt = int(round(T/dt))    # no of intervals

    outfile = open(filename, 'w')
    # u: time level n+1, u_1: time level n
    t = 0
    u_1 = I
    outfile.write('%16E %16E\n' % (t, u_1))
    for n in range(1, Nt+1):
        u = (1 - (1-theta)*a*dt)/(1 + theta*dt*a)*u_1
        u_1 = u
        t += dt
        outfile.write('%16E %16E\n' % (t, u))
    outfile.close()
    return u, t
```

This code snippet serves as a quick introduction to file writing in Python. Reading the data in the file into arrays `t` and `u` are done by the function

```
def read_file(filename='sol.dat'):
    infile = open(filename, 'r')
    u = []; t = []
    for line in infile:
        words = line.split()
        if len(words) != 2:
            print 'Found more than two numbers on a line!', words
            sys.exit(1) # abort
        t.append(float(words[0]))
        u.append(float(words[1]))
    return np.array(t), np.array(u)
```

This type of file with numbers in rows and columns is very common, and `numpy` has a function `loadtxt` which loads such tabular data into a two-dimensional array, say with name `data`. The number in row `i` and column `j` is then `data[i,j]`. The whole column number `j` can be extracted by `data[:,j]`. A version of `read_file` using `np.loadtxt` reads

```
def read_file_numpy(filename='sol.dat'):
    data = np.loadtxt(filename)
    t = data[:,0]
    u = data[:,1]
    return t, u
```

The present counterpart to the `explore` function from `decay_plot_mpl.py` must run `solver_memsave` and then load data from file before we can compute the error measure and make the plot:

```
def explore(I, a, T, dt, theta=0.5, makeplot=True):
    filename = 'u.dat'
    u, t = solver_memsave(I, a, T, dt, theta, filename)

    t, u = read_file(filename)
    u_e = exact_solution(t, I, a)
    e = u_e - u
    E = sqrt(dt*np.sum(e**2))
    if makeplot:
        figure()
    ...
```

Apart from the internal implementation, where  $u^n$  values are stored in a file rather than in an array, `decay_memsave.py` file works exactly as the `decay_plot_mpl.py` file.

### 3 Exercises

#### Exercise 1: Differentiate a function

Given a discrete mesh function  $u^n$  as an arrays `u` with  $u^n$  values at mesh points  $t_n = n\Delta t$ , write a function `differentiate(u, dt)` that returns the discrete derivative  $d^n$  of the mesh function  $u^n$ . The parameter `dt` reflects the mesh spacing  $\Delta t$ . The discrete derivative can be based on centered differences:

$$d^n = \frac{u^{n+1} - u^{n-1}}{2\Delta t}, \quad n = 1, \dots, N_t - 1.$$

At the end points we use forward and backward differences:

$$d^0 = \frac{u^1 - u^0}{\Delta t},$$

and

$$d^{N_t} = \frac{u^{N_t} - u^{N_t-1}}{\Delta t}.$$

**Hint.** The three differentiation formulas are exact for quadratic polynomials. Use this property to verify the program.  
Filename: `differentiate.py`.

## Exercise 2: Experiment with integer division

Explain what happens in the following computations, where some are mathematically unexpected:

```
>>> dt = 3
>>> T = 8
>>> Nt = T/dt
>>> Nt
2
>>> theta = 1; a = 1
>>> (1 - (1-theta)*a*dt)/(1 + theta*dt*a)
0
```

Filename: pyproblems.txt.

## Exercise 3: Experiment with wrong computations

Consider the `solver` function in the `decay_v1.py` file and the following call:

```
u, t = solver(I=1, a=1, T=7, dt=2, theta=1)
```

The output becomes

```
t= 0.000 u=1
t= 2.000 u=0
t= 4.000 u=0
t= 6.000 u=0
```

Print out the result of all intermediate computations and use `type(v)` to see the object type of the result stored in `v`. Examine the intermediate calculations and explain why `u` is wrong and why we compute up to  $t = 6$  only even though we specified  $T = 7$ . Filename: `decay_v1_err.py`.

## Exercise 4: Plot the error function

Solve the problem  $u' = -au$ ,  $u(0) = I$ , using the Forward Euler, Backward Euler, and Crank-Nicolson schemes. For each scheme, plot the error mesh function  $e^n = u_e(t_n) - u^n$  for  $\Delta t$ ,  $\frac{1}{4}\Delta t$ , and  $\frac{1}{8}\Delta t$ , where  $u_e$  is the exact solution of the ODE and  $u^n$  is the numerical solution at mesh point  $t_n$ . Filename: `decay_plot_error.py`.

## Exercise 5: Change formatting of numbers and debug

The `decay_memsave.py` program writes the time values and solution values to a file which looks like

```
0.0000000000000000E+00 1.0000000000000000E+00
2.0000000000000001E-01 8.333333333333337E-01
4.0000000000000002E-01 6.944444444444445E-01
6.0000000000000009E-01 5.787037037037038E-01
8.0000000000000004E-01 4.822530864197532E-01
1.0000000000000000E+00 4.018775720164610E-01
1.2000000000000000E+00 3.348979766803841E-01
1.3999999999999999E+00 2.790816472336534E-01
```

Modify the file output such that it looks like

```
0.000 1.00000
0.200 0.83333
0.400 0.69444
0.600 0.57870
0.800 0.48225
1.000 0.40188
1.200 0.33490
1.400 0.27908
```

If you have just modified the formatting of numbers in the file, running the modified program

---

```
Terminal> python decay_memsave_v2.py --T 10 --theta 1 \
--dt 0.2 --makeplot
```

---

leads to Bug in the implementation!. Why? Filename: decay\_memsave\_v2.py.

## 4 Analysis of finite difference equations

We address the ODE for exponential decay,

$$u'(t) = -au(t), \quad u(0) = I, \quad (57)$$

where  $a$  and  $I$  are given constants. This problem is solved by the  $\theta$ -rule finite difference scheme, resulting in the recursive equations

$$u^{n+1} = \frac{1 - (1 - \theta)a\Delta t}{1 + \theta a\Delta t} u^n \quad (58)$$

for the numerical solution  $u^{n+1}$ , which approximates the exact solution  $u_e$  at time point  $t_{n+1}$ . For constant mesh spacing, which we assume here,  $t_{n+1} = (n+1)\Delta t$ .

**Discouraging numerical solutions.** Choosing  $I = 1$ ,  $a = 2$ , and running experiments with  $\theta = 1, 0.5, 0$  for  $\Delta t = 1.25, 0.75, 0.5, 0.1$ , gives the results in Figures 10, 11, and 12.

The characteristics of the displayed curves can be summarized as follows:

- The Backward Euler scheme always gives a monotone solution, lying above the exact curve.
- The Crank-Nicolson scheme gives the most accurate results, but for  $\Delta t = 1.25$  the solution oscillates.
- The Forward Euler scheme gives a growing, oscillating solution for  $\Delta t = 1.25$ ; a decaying, oscillating solution for  $\Delta t = 0.75$ ; a strange solution  $u^n = 0$  for  $n \geq 1$  when  $\Delta t = 0.5$ ; and a solution seemingly as accurate as the one by the Backward Euler scheme for  $\Delta t = 0.1$ , but the curve lies below the exact solution.

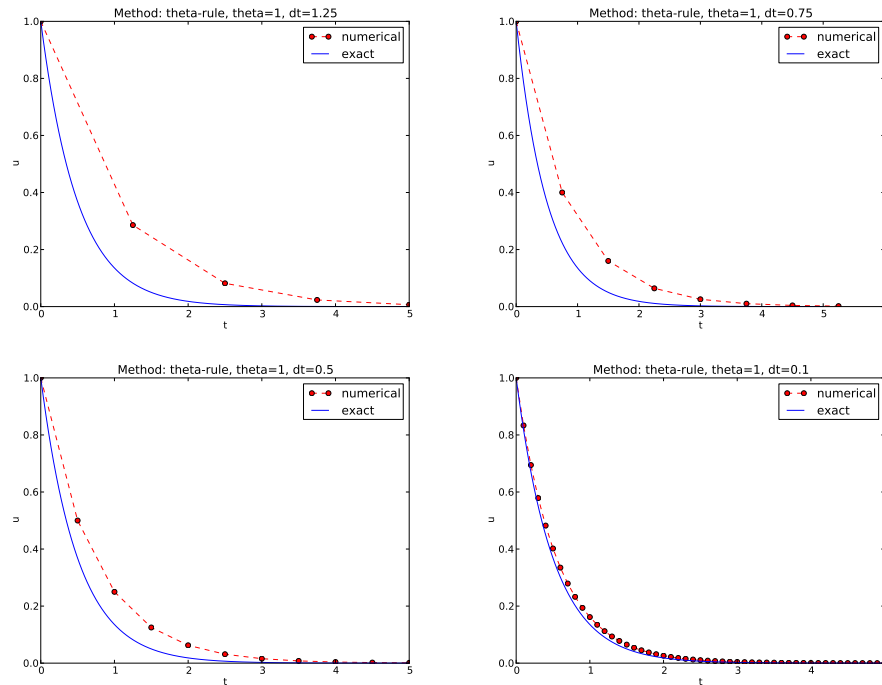


Figure 10: Backward Euler.

Since the exact solution of our model problem is a monotone function,  $u(t) = Ie^{-at}$ , some of these qualitatively wrong results are indeed alarming!

### Goal.

We ask the question

- Under what circumstances, i.e., values of the input data  $I$ ,  $a$ , and  $\Delta t$  will the Forward Euler and Crank-Nicolson schemes result in undesired oscillatory solutions?

The question will be investigated both by numerical experiments and by precise mathematical theory. The latter will help establish general criteria on  $\Delta t$  for avoiding non-physical oscillatory or growing solutions.

Another question to be raised is

- How does  $\Delta t$  impact the error in the numerical solution?



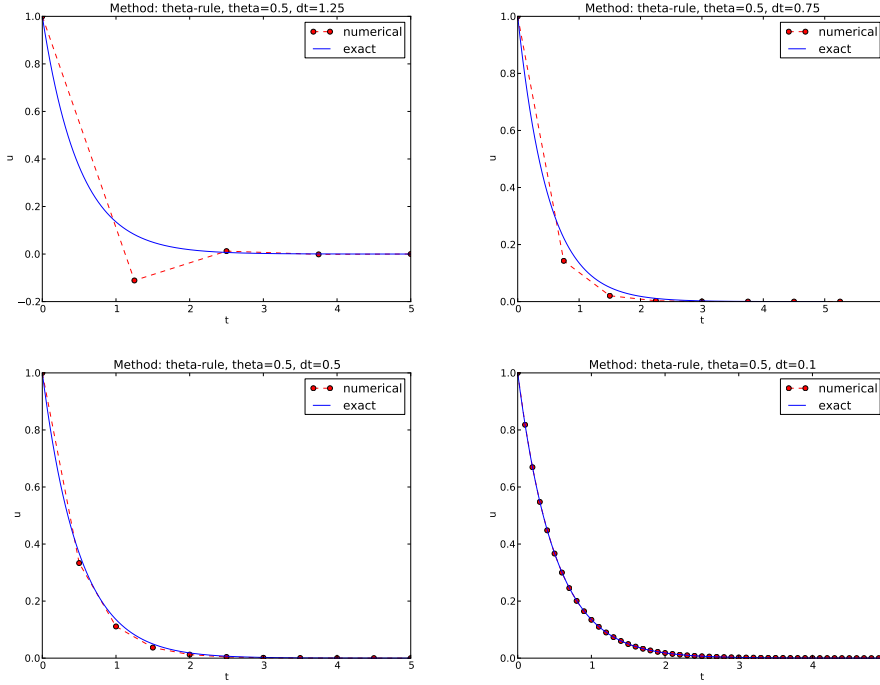


Figure 11: Crank-Nicolson.

For our simple model problem we can answer this question very precisely, but we will also look at simplified formulas for small  $\Delta t$  and touch upon important concepts such as *convergence rate* and *the order of a scheme*. Other fundamental concepts mentioned are stability, consistency, and convergence.

#### 4.1 Experimental investigation of oscillatory solutions

To address the first question above, we may set up an experiment where we loop over values of  $I$ ,  $a$ , and  $\Delta t$ . For each experiment, we flag the solution as oscillatory if

$$u^n > u^{n-1},$$

for some value of  $n$ , since we expect  $u^n$  to decay with  $n$ , but oscillations make  $u$  increase over a time step. We will quickly see that oscillations are independent of  $I$ , but do depend on  $a$  and  $\Delta t$ . Therefore, we introduce a two-dimensional function  $B(a, \Delta t)$  which is 1 if oscillations occur and 0 otherwise. We can visualize  $B$  as a contour plot (lines for which  $B = \text{const}$ ). The contour  $B = 0.5$

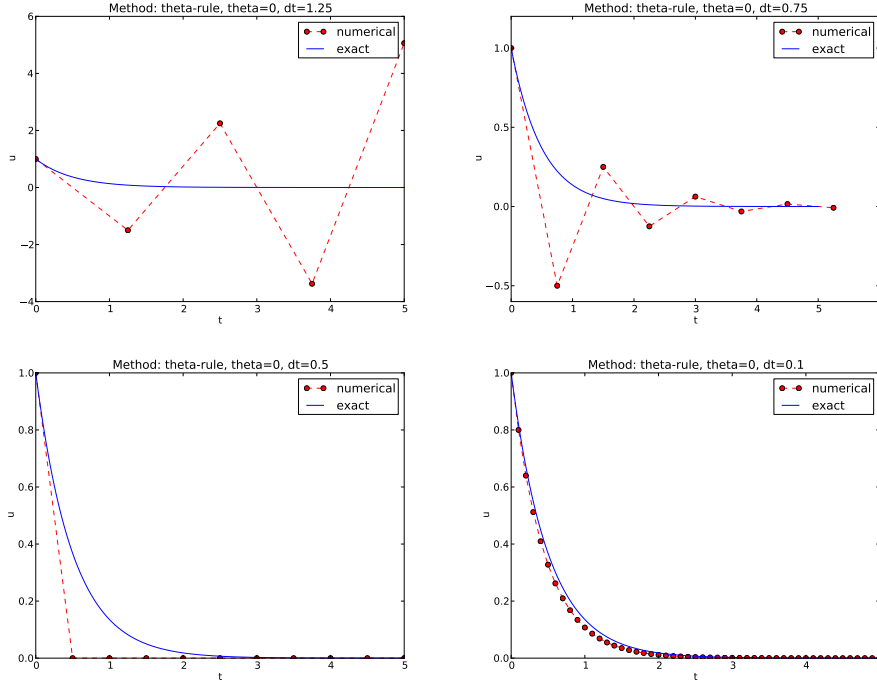


Figure 12: Forward Euler.

corresponds to the borderline between oscillatory regions with  $B = 1$  and monotone regions with  $B = 0$  in the  $a, \Delta t$  plane.

The  $B$  function is defined at discrete  $a$  and  $\Delta t$  values. Say we have given  $P$   $a$  values,  $a_0, \dots, a_{P-1}$ , and  $Q$   $\Delta t$  values,  $\Delta t_0, \dots, \Delta t_{Q-1}$ . These  $a_i$  and  $\Delta t_j$  values,  $i = 0, \dots, P-1$ ,  $j = 0, \dots, Q-1$ , form a rectangular mesh of  $P \times Q$  points in the plane. At each point  $(a_i, \Delta t_j)$ , we associate the corresponding value of  $B(a_i, \Delta t_j)$ , denoted  $B_{ij}$ . The  $B_{ij}$  values are naturally stored in a two-dimensional array. We can thereafter create a plot of the contour line  $B_{ij} = 0.5$  dividing the oscillatory and monotone regions. The file `decay_osc_regions.py` (osc\_regions stands for "oscillatory regions") contains all nuts and bolts to produce the  $B = 0.5$  line in Figures 13 and 14. The oscillatory region is above this line.

```
from decay_mod import solver
import numpy as np
import scitools.std as st

def non_physical_behavior(I, a, T, dt, theta):
    """
    Given lists/arrays a and dt, and numbers I, dt, and theta,
    make a two-dimensional contour line B=0.5, where B=1>0.5
    means oscillatory (unstable) solution, and B=0<0.5 means
    monotone solution of u'=-au.
```

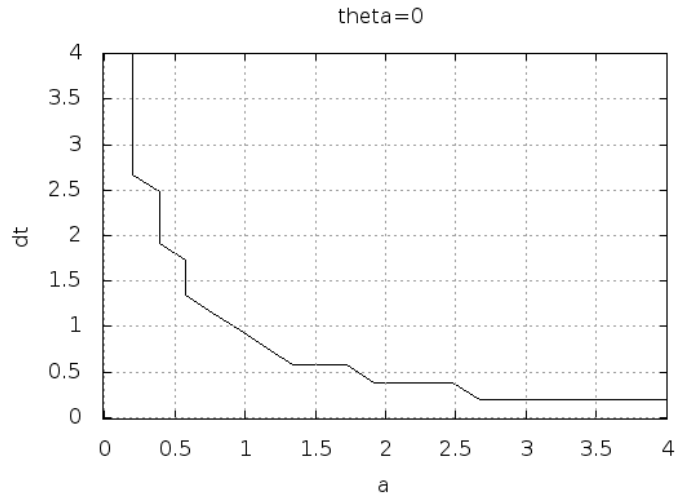


Figure 13: Forward Euler scheme: oscillatory solutions occur for points above the curve.

```
"""
a = np.asarray(a); dt = np.asarray(dt) # must be arrays
B = np.zeros((len(a), len(dt)))        # results
for i in range(len(a)):
    for j in range(len(dt)):
        u, t = solver(I, a[i], T, dt[j], theta)
        # Does u have the right monotone decay properties?
        correct_qualitative_behavior = True
        for n in range(1, len(u)):
            if u[n] > u[n-1]: # Not decaying?
                correct_qualitative_behavior = False
                break # Jump out of loop
        B[i,j] = float(correct_qualitative_behavior)
a_, dt_ = st.ndgrid(a, dt) # make mesh of a and dt values
st.contour(a_, dt_, B, 1)
st.grid('on')
st.title('theta=%g' % theta)
st.xlabel('a'); st.ylabel('dt')
st.savefig('osc_region_theta_%s.png' % theta)
st.savefig('osc_region_theta_%s.pdf' % theta)

non_physical_behavior(
    I=1,
    a=np.linspace(0.01, 4, 22),
    dt=np.linspace(0.01, 4, 22),
    T=6,
    theta=0.5)
```

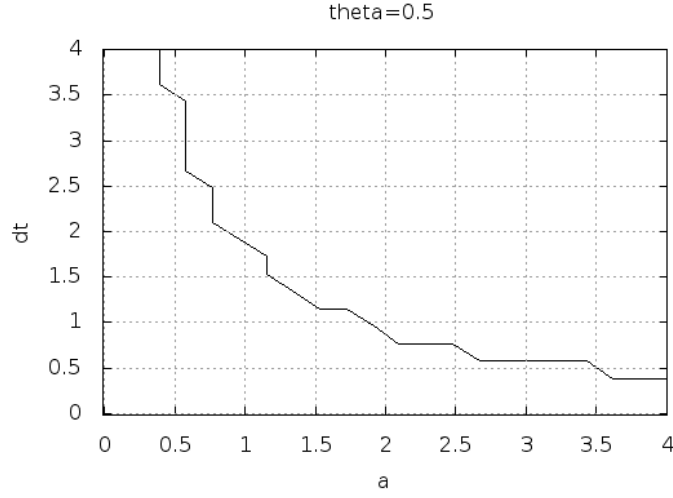


Figure 14: Crank-Nicolson scheme: oscillatory solutions occur for points above the curve.

By looking at the curves in the figures one may guess that  $a\Delta t$  must be less than a critical limit to avoid the undesired oscillations. This limit seems to be about 2 for Crank-Nicolson and 1 for Forward Euler. We shall now establish a precise mathematical analysis of the discrete model that can explain the observations in our numerical experiments.

## 4.2 Exact numerical solution

Starting with  $u^0 = I$ , the simple recursion (58) can be applied repeatedly  $n$  times, with the result that

$$u^n = IA^n, \quad A = \frac{1 - (1 - \theta)a\Delta t}{1 + \theta a\Delta t}. \quad (59)$$

### Solving difference equations.

Difference equations where all terms are linear in  $u^{n+1}$ ,  $u^n$ , and maybe  $u^{n-1}$ ,  $u^{n-2}$ , etc., are called *homogeneous*, *linear* difference equations, and their solutions are generally of the form  $u^n = A^n$ . Inserting this expression and dividing by  $A^{n+1}$  gives a polynomial equation in  $A$ . In the present

case we get

$$A = \frac{1 - (1 - \theta)a\Delta t}{1 + \theta a\Delta t}.$$

This is a solution technique of wider applicability than repeated use of the recursion (58).

Regardless of the solution approach, we have obtained a formula for  $u^n$ . This formula can explain everything what we see in the figures above, but it also gives us a more general insight into accuracy and stability properties of the three schemes.

### 4.3 Stability

Since  $u^n$  is a factor  $A$  raised to an integer power  $n$ , we realize that  $A < 0$  will for odd powers imply  $u^n < 0$  and for even power result in  $u^n > 0$ . That is, the solution oscillates between the mesh points. We have oscillations due to  $A < 0$  when

$$(1 - \theta)a\Delta t > 1. \quad (60)$$

Since  $A > 0$  is a requirement for having a numerical solution with the same basic property (monotonicity) as the exact solution, we may say that  $A > 0$  is a *stability criterion*. Expressed in terms of  $\Delta t$  the stability criterion reads

$$\Delta t < \frac{1}{(1 - \theta)a}. \quad (61)$$

The Backward Euler scheme is always stable since  $A < 0$  is impossible for  $\theta = 1$ , while non-oscillating solutions for Forward Euler and Crank-Nicolson demand  $\Delta t \leq 1/a$  and  $\Delta t \leq 2/a$ , respectively. The relation between  $\Delta t$  and  $a$  look reasonable: a larger  $a$  means faster decay and hence a need for smaller time steps.

Looking at Figure 12, we see that with  $a\Delta t = 2 \cdot 1.25 = 2.5$ ,  $A = -1.5$ , and the solution  $u^n = (-1.5)^n$  oscillates *and* grows. With  $a\Delta t = 2 \cdot 0.75 = 1.5$ ,  $A = -0.5$ ,  $u^n = (-0.5)^n$  decays but oscillates. The peculiar case  $\Delta t = 0.5$ , where the Forward Euler scheme produces a solution that is stuck on the  $t$  axis, corresponds to  $A = 0$  and therefore  $u^0 = I = 1$  and  $u^n = 0$  for  $n \geq 1$ . The decaying oscillations in the Crank-Nicolson scheme for  $\Delta t = 1.25$  are easily explained by the fact that  $A \approx -0.11 < 0$ .

The factor  $A$  is called the *amplification factor* since the solution at a new time level is  $A$  times the solution at the previous time level. For a decay process, we must obviously have  $|A| \leq 1$ , which is fulfilled for all  $\Delta t$  if  $\theta \geq 1/2$ . Arbitrarily large values of  $u$  can be generated when  $|A| > 1$  and  $n$  is large enough. The numerical solution is in such cases totally irrelevant to an ODE modeling decay processes! To avoid this situation, we must for  $\theta < 1/2$  have

$$\Delta t \leq \frac{2}{(1 - 2\theta)a}, \quad (62)$$

which means  $\Delta t < 2/a$  for the Forward Euler scheme.

#### Stability properties.

We may summarize the stability investigations as follows:

1. The Forward Euler method is a *conditionally stable* scheme because it requires  $\Delta t < 2/a$  for avoiding growing solutions and  $\Delta t < 1/a$  for avoiding oscillatory solutions.
2. The Crank-Nicolson is *unconditionally stable* with respect to growing solutions, while it is conditionally stable with the criterion  $\Delta t < 2/a$  for avoiding oscillatory solutions.
3. The Backward Euler method is unconditionally stable with respect to growing and oscillatory solutions - any  $\Delta t$  will work.

Much literature on ODEs speaks about L-stable and A-stable methods. In our case A-stable methods ensures non-growing solutions, while L-stable methods also avoids oscillatory solutions.

## 4.4 Comparing amplification factors

After establishing how  $A$  impacts the qualitative features of the solution, we shall now look more into how well the numerical amplification factor approximates the exact one. The exact solution reads  $u(t) = Ie^{-at}$ , which can be rewritten as

$$u_e(t_n) = Ie^{-an\Delta t} = I(e^{-a\Delta t})^n. \quad (63)$$

From this formula we see that the exact amplification factor is

$$A_e = e^{-a\Delta t}. \quad (64)$$

We realize that the exact and numerical amplification factors depend on  $a$  and  $\Delta t$  through the product  $a\Delta t$ . Therefore, it is convenient to introduce a symbol for this product,  $p = a\Delta t$ , and view  $A$  and  $A_e$  as functions of  $p$ . Figure 15 shows these functions. Crank-Nicolson is clearly closest to the exact amplification factor, but that method has the unfortunate oscillatory behavior when  $p > 2$ .

## 4.5 Series expansion of amplification factors

As an alternative to the visual understanding inherent in Figure 15, there is a strong tradition in numerical analysis to establish formulas for the approximation

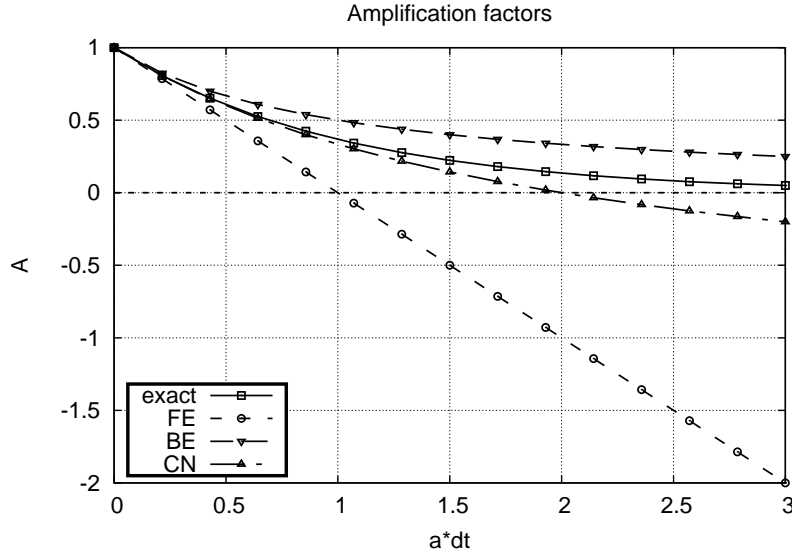


Figure 15: Comparison of amplification factors.

errors when the discretization parameter, here  $\Delta t$ , becomes small. In the present case we let  $p$  be our small discretization parameter, and it makes sense to simplify the expressions for  $A$  and  $A_e$  by using Taylor polynomials around  $p = 0$ . The Taylor polynomials are accurate for small  $p$  and greatly simplifies the comparison of the analytical expressions since we then can compare polynomials, term by term.

Calculating the Taylor series for  $A_e$  is easily done by hand, but the three versions of  $A$  for  $\theta = 0, 1, \frac{1}{2}$  lead to more cumbersome calculations. Nowadays, analytical computations can benefit greatly by symbolic computer algebra software. The Python package `sympy` represents a powerful computer algebra system, not yet as sophisticated as the famous Maple and Mathematica systems, but free and very easy to integrate with our numerical computations in Python.

When using `sympy`, it is convenient to enter the interactive Python mode where we can write expressions and statements and immediately see the results. Here is a simple example. We strongly recommend to use `isympy` (or `ipython`) for such interactive sessions.

Let us illustrate `sympy` with a standard Python shell syntax (`>>>` prompt) to compute a Taylor polynomial approximation to  $e^{-p}$ :

```
>>> from sympy import *
>>> # Create p as a mathematical symbol with name 'p'
>>> p = Symbol('p')
>>> # Create a mathematical expression with p
>>> A_e = exp(-p)
>>>
>>> # Find the first 6 terms of the Taylor series of A_e
```

```
>>> A_e.series(p, 0, 6)
1 + (1/2)*p**2 - p - 1/6*p**3 - 1/120*p**5 + (1/24)*p**4 + O(p**6)
```

Lines with >>> represent input lines and lines without this prompt represents the result of computations (note that `isympy` and `ipython` apply other prompts, but in this text we always apply >>> for interactive Python computing). Apart from the order of the powers, the computed formula is easily recognized as the beginning of the Taylor series for  $e^{-p}$ .

Let us define the numerical amplification factor where  $p$  and  $\theta$  enter the formula as symbols:

```
>>> theta = Symbol('theta')
>>> A = (1-(1-theta)*p)/(1+theta*p)
```

To work with the factor for the Backward Euler scheme we can substitute the value 1 for `theta`:

```
>>> A.subs(theta, 1)
1/(1 + p)
```

Similarly, we can replace `theta` by  $1/2$  for Crank-Nicolson, preferably using an exact rational representation of  $1/2$  in `sympy`:

```
>>> half = Rational(1,2)
>>> A.subs(theta, half)
1/(1 + (1/2)*p)*(1 - 1/2*p)
```

The Taylor series of the amplification factor for the Crank-Nicolson scheme can be computed as

```
>>> A.subs(theta, half).series(p, 0, 4)
1 + (1/2)*p**2 - p - 1/4*p**3 + O(p**4)
```

We are now in a position to compare Taylor series:

```
>>> FE = A_e.series(p, 0, 4) - A.subs(theta, 0).series(p, 0, 4)
>>> BE = A_e.series(p, 0, 4) - A.subs(theta, 1).series(p, 0, 4)
>>> CN = A_e.series(p, 0, 4) - A.subs(theta, half).series(p, 0, 4)
>>> FE
(1/2)*p**2 - 1/6*p**3 + O(p**4)
>>> BE
-1/2*p**2 + (5/6)*p**3 + O(p**4)
>>> CN
(1/12)*p**3 + O(p**4)
```

From these expressions we see that the error  $A - A_e \sim \mathcal{O}(p^2)$  for the Forward and Backward Euler schemes, while  $A - A_e \sim \mathcal{O}(p^3)$  for the Crank-Nicolson scheme. It is the *leading order term*, i.e., the term of the lowest order (polynomial degree), that is of interest, because as  $p \rightarrow 0$ , this term is (much) bigger than the higher-order terms (think of  $p = 0.01$ :  $p$  is a hundred times larger than  $p^2$ ).



Now,  $a$  is a given parameter in the problem, while  $\Delta t$  is what we can vary. One therefore usually writes the error expressions in terms  $\Delta t$ . When then have

$$A - A_e = \begin{cases} \mathcal{O}(\Delta t^2), & \text{Forward and Backward Euler,} \\ \mathcal{O}(\Delta t^3), & \text{Crank-Nicolson} \end{cases} \quad (65)$$

We say that the Crank-Nicolson scheme has an error in the amplification factor of order  $\Delta t^3$ , while the two other schemes are of order  $\Delta t^2$  in the same quantity. What is the significance of the order expression? If we halve  $\Delta t$ , the error in amplification factor at a time level will be reduced by a factor of 4 in the Forward and Backward Euler schemes, and by a factor of 8 in the Crank-Nicolson scheme. That is, as we reduce  $\Delta t$  to obtain more accurate results, the Crank-Nicolson scheme reduces the error more efficiently than the other schemes.

#### 4.6 The fraction of numerical and exact amplification factors

An alternative comparison of the schemes is to look at the ratio  $A/A_e$ , or the error  $1 - A/A_e$  in this ratio:

```
>>> FE = 1 - (A.subs(theta, 0)/A_e).series(p, 0, 4)
>>> BE = 1 - (A.subs(theta, 1)/A_e).series(p, 0, 4)
>>> CN = 1 - (A.subs(theta, half)/A_e).series(p, 0, 4)
>>> FE
(1/2)*p**2 + (1/3)*p**3 + O(p**4)
>>> BE
-1/2*p**2 + (1/3)*p**3 + O(p**4)
>>> CN
(1/12)*p**3 + O(p**4)
```

The leading-order terms have the same powers as in the analysis of  $A - A_e$ .

#### 4.7 The global error at a point

The error in the amplification factor reflects the error when progressing from time level  $t_n$  to  $t_{n-1}$ . To investigate the real error at a point, known as the *global error*, we look at  $e^n = u^n - u_e(t_n)$  for some  $n$  and Taylor expand the mathematical expressions as functions of  $p = a\Delta t$ :

```
>>> n = Symbol('n')
>>> u_e = exp(-p*n)
>>> u_n = A**n
>>> FE = u_e.series(p, 0, 4) - u_n.subs(theta, 0).series(p, 0, 4)
>>> BE = u_e.series(p, 0, 4) - u_n.subs(theta, 1).series(p, 0, 4)
>>> CN = u_e.series(p, 0, 4) - u_n.subs(theta, half).series(p, 0, 4)
>>> FE
(1/2)*n*p**2 - 1/2*n**2*p**3 + (1/3)*n*p**3 + O(p**4)
>>> BE
(1/2)*n**2*p**3 - 1/2*n*p**2 + (1/3)*n*p**3 + O(p**4)
>>> CN
(1/12)*n*p**3 + O(p**4)
```

For a fixed time  $t$ , the parameter  $n$  in these expressions increases as  $p \rightarrow 0$  since  $t = n\Delta t = \text{const}$  and hence  $n$  must increase like  $\Delta t^{-1}$ . With  $n$  substituted by  $t/\Delta t$  in the leading-order error terms, these become  $\frac{1}{2}na^2\Delta t^2 = \frac{1}{2}ta^2\Delta t$  for the Forward and Backward Euler scheme, and  $\frac{1}{12}na^3\Delta t^3 = \frac{1}{12}ta^3\Delta t^2$  for the Crank-Nicolson scheme. The global error is therefore of second order (in  $\Delta t$ ) for the latter scheme and of first order for the former schemes.

When the global error  $e^n \rightarrow 0$  as  $\Delta t \rightarrow 0$ , we say that the scheme is *convergent*. It means that the numerical solution approaches the exact solution as the mesh is refined, and this is a much desired property of a numerical method.

## 4.8 Integrated errors

It is common to study the norm of the numerical error, as explained in detail in Section 2.4. The  $L^2$  norm can be computed by treating  $e^n$  as a function of  $t$  in *sympy* and performing symbolic integration. For the Forward Euler scheme we have

```
p, n, a, dt, t, T, theta = symbols('p n a dt t T 'theta')
A = (1-(1-theta)*p)/(1+theta*p)
u_e = exp(-p*n)
u_n = A**n
error = u_e.series(p, 0, 4) - u_n.subs(theta, 0).series(p, 0, 4)
# Introduce t and dt instead of n and p
error = error.subs('n', 't/dt').subs(p, 'a*dt')
error = error.as_leading_term(dt) # study only the first term
print error
error_L2 = sqrt(integrate(error**2, (t, 0, T)))
print error_L2
```

The output reads

$$\sqrt{30} * \sqrt{T^3 a^4 dt^2 (6 T^2 a^2 - 15 T a + 10)} / 60$$

which means that the  $L^2$  error behaves like  $a^2 \Delta t$ .

Strictly speaking, the numerical error is only defined at the mesh points so it makes most sense to compute the  $\ell^2$  error

$$\|e^n\|_{\ell^2} = \sqrt{\Delta t \sum_{n=0}^{N_t} (u_e(t_n) - u^n)^2}.$$

We have obtained an exact analytical expressions for the error at  $t = t_n$ , but here we use the leading-order error term only since we are mostly interested in how the error behaves as a polynomial in  $\Delta t$ , and then the leading order term will dominate. For the Forward Euler scheme,  $u_e(t_n) - u^n \approx \frac{1}{2}np^2$ , and we have

$$\|e^n\|_{\ell^2}^2 = \Delta t \sum_{n=0}^{N_t} \frac{1}{4} n^2 p^4 = \Delta t \frac{1}{4} p^4 \sum_{n=0}^{N_t} n^2.$$

Now,  $\sum_{n=0}^{N_t} n^2 \approx \frac{1}{3} N_t^3$ . Using this approximation, setting  $N_t = T/\Delta t$ , and taking the square root gives the expression

$$\|e^n\|_{\ell^2} = \frac{1}{2} \sqrt{\frac{T^3}{3}} a^2 \Delta t.$$

Calculations for the Backward Euler scheme are very similar and provide the same result, while the Crank-Nicolson scheme leads to

$$\|e^n\|_{\ell^2} = \frac{1}{12} \sqrt{\frac{T^3}{3}} a^3 \Delta t^2.$$

#### Summary of errors.

Both the point-wise and the time-integrated true errors are of second order in  $\Delta t$  for the Crank-Nicolson scheme and of first order in  $\Delta t$  for the Forward Euler and Backward Euler schemes.

## 4.9 Truncation error

The truncation error is a very frequently used error measure for finite difference methods. It is defined as *the error in the difference equation that arises when inserting the exact solution*. Contrary to many other error measures, e.g., the true error  $e^n = u_e(t_n) - u^n$ , the truncation error is a quantity that is easily computable.

Before reading on, it is wise to review Section 1.7 on how Taylor polynomials were used to derive finite differences and quantify the error in the formulas. Very similar reasoning, and almost identical mathematical details, will be carried out below, but in a slightly different context. Here, our focus is on the error when solving a differential equation, not only in the formula of a finite difference, but these errors are tightly connected.

Let us illustrate the calculation of the truncation error for the Forward Euler scheme. We start with the difference equation on operator form,

$$[D_t u = -au]^n,$$

i.e.,

$$\frac{u^{n+1} - u^n}{\Delta t} = -au^n.$$

The idea is to see how well the exact solution  $u_e(t)$  fulfills this equation. Since  $u_e(t)$  in general will not obey the discrete equation, error in the discrete equation, called a *residual*, denoted here by  $R^n$ :

$$R^n = \frac{u_e(t_{n+1}) - u_e(t_n)}{\Delta t} + au_e(t_n). \quad (66)$$

The residual is defined at each mesh point and is therefore a mesh function with a superscript  $n$ .

The interesting feature of  $R^n$  is to see how it depends on the discretization parameter  $\Delta t$ . The tool for reaching this goal is to Taylor expand  $u_e$  around the point where the difference equation is supposed to hold, here  $t = t_n$ . We have that

$$u_e(t_{n+1}) = u_e(t_n) + u'_e(t_n)\Delta t + \frac{1}{2}u''_e(t_n)\Delta t^2 + \dots$$

Inserting this Taylor series in (66) gives

$$R^n = u'_e(t_n) + \frac{1}{2}u''_e(t_n)\Delta t + \dots + au_e(t_n).$$

Now,  $u_e$  fulfills the ODE  $u'_e = -au_e$  such that the first and last term cancels and we have

$$R^n \approx \frac{1}{2}u''_e(t_n)\Delta t.$$

This  $R^n$  is the *truncation error*, which for the Forward Euler is seen to be of first order in  $\Delta t$ .

The above procedure can be repeated for the Backward Euler and the Crank-Nicolson schemes. We start with the scheme in operator notation, write it out in detail, Taylor expand  $u_e$  around the point  $\tilde{t}$  at which the difference equation is defined, collect terms that correspond to the ODE (here  $u'_e + au_e$ ), and identify the remaining terms as the residual  $R$ , which is the truncation error. The Backward Euler scheme leads to

$$R^n \approx -\frac{1}{2}u''_e(t_n)\Delta t,$$

while the Crank-Nicolson scheme gives

$$R^{n+\frac{1}{2}} \approx \frac{1}{24}u'''_e(t_{n+\frac{1}{2}})\Delta t^2.$$

The *order*  $r$  of a finite difference scheme is often defined through the leading term  $\Delta t^r$  in the truncation error. The above expressions point out that the Forward and Backward Euler schemes are of first order, while Crank-Nicolson is of second order. We have looked at other error measures in other sections, like the error in amplification factor and the error  $e^n = u_e(t_n) - u^n$ , and expressed these error measures in terms of  $\Delta t$  to see the order of the method. Normally, calculating the truncation error is more straightforward than deriving the expressions for other error measures and therefore the easiest way to establish the order of a scheme.

#### 4.10 Consistency, stability, and convergence

Three fundamental concepts when solving differential equations by numerical methods are consistency, stability, and convergence. We shall briefly touch these concepts below in the context of the present model problem.

Consistency means that the error in the difference equation, measured through the truncation error, goes to zero as  $\Delta t \rightarrow 0$ . Since the truncation error tells how well the exact solution fulfills the difference equation, and the exact solution fulfills the differential equation, consistency ensures that the difference equation approaches the differential equation in the limit. The expressions for the truncation errors in the previous section are all proportional to  $\Delta t$  or  $\Delta t^2$ , hence they vanish as  $\Delta t \rightarrow 0$ , and all the schemes are consistent. Lack of consistency implies that we actually solve a different differential equation in the limit  $\Delta t \rightarrow 0$  than we aim at.

Stability means that the numerical solution exhibits the same qualitative properties as the exact solution. This is obviously a feature we want the numerical solution to have. In the present exponential decay model, the exact solution is monotone and decaying. An increasing numerical solution is not in accordance with the decaying nature of the exact solution and hence unstable. We can also say that an oscillating numerical solution lacks the property of monotonicity of the exact solution and is also unstable. We have seen that the Backward Euler scheme always leads to monotone and decaying solutions, regardless of  $\Delta t$ , and is hence stable. The Forward Euler scheme can lead to increasing solutions and oscillating solutions if  $\Delta t$  is too large and is therefore unstable unless  $\Delta t$  is sufficiently small. The Crank-Nicolson can never lead to increasing solutions and has no problem to fulfill that stability property, but it can produce oscillating solutions and is unstable in that sense, unless  $\Delta t$  is sufficiently small.

Convergence implies that the global (true) error mesh function  $e^n = u_e(t_n) - u^n \rightarrow 0$  as  $\Delta t \rightarrow 0$ . This is really what we want: the numerical solution gets as close to the exact solution as we request by having a sufficiently fine mesh.

Convergence is hard to establish theoretically, except in quite simple problems like the present one. Stability and consistency are much easier to calculate. A major breakthrough in the understanding of numerical methods for differential equations came in 1956 when Lax and Richtmeyer established equivalence between convergence on one hand and consistency and stability on the other (the [Lax equivalence theorem](#)). In practice it meant that one can first establish that a method is stable and consistent, and then it is automatically convergent (which is much harder to establish). The result holds for linear problems only, and in the world of nonlinear differential equations the relations between consistency, stability, and convergence are much more complicated.

We have seen in the previous analysis that the Forward Euler, Backward Euler, and Crank-Nicolson schemes are convergent ( $e^n \rightarrow 0$ ), that they are consistent ( $R^n \rightarrow 0$ ), and that they are stable under certain conditions on the size of  $\Delta t$ . We have also derived explicit mathematical expressions for  $e^n$ , the truncation error, and the stability criteria.

## 5 Exercises

### Exercise 6: Visualize the accuracy of finite differences

The purpose of this exercise is to visualize the accuracy of finite difference approximations of the derivative of a given function. For any finite difference approximation, take the Forward Euler difference as an example, and any specific function, take  $u = e^{-at}$ , we may introduce an error fraction

$$E = \frac{[D_t^+ u]^n}{u'(t_n)} = \frac{\exp(-a(t_n + \Delta t)) - \exp(-at_n)}{-a \exp(-at_n) \Delta t} = \frac{1}{a \Delta t} (1 - \exp(-a \Delta t)),$$

and view  $E$  as a function of  $\Delta t$ . We expect that  $\lim_{\Delta t \rightarrow 0} E = 1$ , while  $E$  may deviate significantly from unity for large  $\Delta t$ . How the error depends on  $\Delta t$  is best visualized in a graph where we use a logarithmic scale for  $\Delta t$ , so we can cover many orders of magnitude of that quantity. Here is a code segment creating an array of 100 intervals, on the logarithmic scale, ranging from  $10^{-6}$  to  $10^{-0.5}$  and then plotting  $E$  versus  $p = a \Delta t$  with logarithmic scale on the  $p$  axis:

```
from numpy import logspace, exp
from matplotlib.pyplot import plot, semilogx
p = logspace(-6, -0.5, 101)
y = (1-exp(-p))/p
semilogx(p, y)
```

Illustrate such errors for the finite difference operators  $[D_t^+ u]^n$  (forward),  $[D_t^- u]^n$  (backward), and  $[D_t u]^n$  (centered) in the same plot.

Perform a Taylor series expansions of the error fractions and find the leading order  $r$  in the expressions of type  $1 + Cp^r + \mathcal{O}(p^{r+1})$ , where  $C$  is some constant.

**Hint.** To save manual calculations and learn more about symbolic computing, make functions for the three difference operators and use `sympy` to perform the symbolic differences, differentiation, and Taylor series expansion. To plot a symbolic expression  $E$  against  $p$ , convert the expression to a Python function first: `E = sympy.lamdify([p], E)`.

Filename: `decay_plot_fd_error.py`.

### Exercise 7: Explore the $\theta$ -rule for exponential growth

This exercise asks you to solve the ODE  $u' = -au$  with  $a < 0$  such that the ODE models exponential growth instead of exponential decay. A central theme is to investigate numerical artifacts and non-physical solution behavior.

**a)** Set  $a = -1$  and run experiments with  $\theta = 0, 0.5, 1$  for various values of  $\Delta t$  to uncover numerical artifacts. Recall that the exact solution is a monotone, growing function when  $a < 0$ . Oscillations or significantly wrong growth are signs of wrong qualitative behavior.

From the experiments, select four values of  $\Delta t$  that demonstrate the kind of numerical solutions that are characteristic for this model. Filename: `growth_demo.py`.

b) Write up the amplification factor and plot it for  $\theta = 0, 0.5, 1$  together with the exact one for  $a\Delta t < 0$ . Use the plot to explain the observations made in the experiments.

**Hint.** Modify the `decay_ampf_plot.py` code.  
Filename: `growth_ampf_plot.py`.

## 6 Model extensions

It is time to consider generalizations of the simple decay model  $u = -au$  and also to look at additional numerical solution methods.

### 6.1 Generalization: including a variable coefficient

In the ODE for decay,  $u' = -au$ , we now consider the case where  $a$  depends on time:

$$u'(t) = -a(t)u(t), \quad t \in (0, T], \quad u(0) = I. \quad (67)$$

A Forward Euler scheme consist of evaluating (67) at  $t = t_n$  and approximating the derivative with a forward difference  $[D_t^+ u]^n$ :

$$\frac{u^{n+1} - u^n}{\Delta t} = -a(t_n)u^n. \quad (68)$$

The Backward Euler scheme becomes

$$\frac{u^n - u^{n-1}}{\Delta t} = -a(t_n)u^n. \quad (69)$$

The Crank-Nicolson method builds on sampling the ODE at  $t_{n+\frac{1}{2}}$ . We can evaluate  $a$  at  $t_{n+\frac{1}{2}}$  and use an average for  $u$  at times  $t_n$  and  $t_{n+1}$ :

$$\frac{u^{n+1} - u^n}{\Delta t} = -a(t_{n+\frac{1}{2}})\frac{1}{2}(u^n + u^{n+1}). \quad (70)$$

Alternatively, we can use an average for the product  $au$ :

$$\frac{u^{n+1} - u^n}{\Delta t} = -\frac{1}{2}(a(t_n)u^n + a(t_{n+1})u^{n+1}). \quad (71)$$

The  $\theta$ -rule unifies the three mentioned schemes. One version is to have  $a$  evaluated at  $t_{n+\theta}$ ,

$$\frac{u^{n+1} - u^n}{\Delta t} = -a((1-\theta)t_n + \theta t_{n+1})((1-\theta)u^n + \theta u^{n+1}). \quad (72)$$

Another possibility is to apply a weighted average for the product  $au$ ,

$$\frac{u^{n+1} - u^n}{\Delta t} = -(1-\theta)a(t_n)u^n - \theta a(t_{n+1})u^{n+1}. \quad (73)$$

With the finite difference operator notation the Forward Euler and Backward Euler schemes can be summarized as

$$[D_t^+ u = -au]^n, \quad (74)$$

$$[D_t^- u = -au]^n. \quad (75)$$

The Crank-Nicolson and  $\theta$  schemes depend on whether we evaluate  $a$  at the sample point for the ODE or if we use an average. The various versions are written as

$$[D_t u = -a\bar{u}^t]^{n+\frac{1}{2}}, \quad (76)$$

$$[D_t u = -\bar{a}\bar{u}^t]^{n+\frac{1}{2}}, \quad (77)$$

$$[D_t u = -a\bar{u}^{t,\theta}]^{n+\theta}, \quad (78)$$

$$[D_t u = -\bar{a}\bar{u}^{t,\theta}]^{n+\theta}. \quad (79)$$

## 6.2 Generalization: including a source term

A further extension of the model ODE is to include a source term  $b(t)$ :

$$u'(t) = -a(t)u(t) + b(t), \quad t \in (0, T], \quad u(0) = I. \quad (80)$$

**Schemes.** The time point where we sample the ODE determines where  $b(t)$  is evaluated. For the Crank-Nicolson scheme and the  $\theta$ -rule we have a choice of whether to evaluate  $a(t)$  and  $b(t)$  at the correct point or use an average. The chosen strategy becomes particularly clear if we write up the schemes in the operator notation:

$$[D_t^+ u = -au + b]^n, \quad (81)$$

$$[D_t^- u = -au + b]^n, \quad (82)$$

$$[D_t u = -a\bar{u}^t + b]^{n+\frac{1}{2}}, \quad (83)$$

$$[D_t u = -\bar{a}\bar{u}^t + b]^{n+\frac{1}{2}}, \quad (84)$$

$$[D_t u = -a\bar{u}^{t,\theta} + b]^{n+\theta}, \quad (85)$$

$$[D_t u = -\bar{a}\bar{u}^{t,\theta} + b]^{n+\theta}. \quad (86)$$

## 6.3 Implementation of the generalized model problem

**Deriving the  $\theta$ -rule formula.** Writing out the  $\theta$ -rule in (86), using (43) and (44), we get

$$\frac{u^{n+1} - u^n}{\Delta t} = \theta(-a^{n+1}u^{n+1} + b^{n+1}) + (1 - \theta)(-a^n u^n + b^n), \quad (87)$$



where  $a^n$  means evaluating  $a$  at  $t = t_n$  and similar for  $a^{n+1}$ ,  $b^n$ , and  $b^{n+1}$ . We solve for  $u^{n+1}$ :

$$u^{n+1} = ((1 - \Delta t(1 - \theta)a^n)u^n + \Delta t(\theta b^{n+1} + (1 - \theta)b^n))(1 + \Delta t\theta a^{n+1})^{-1}. \quad (88)$$

**The Python code.** Here is a suitable implementation of (87) where  $a(t)$  and  $b(t)$  are given as Python functions:

```
def solver(I, a, b, T, dt, theta):
    """
    Solve u'=-a(t)*u + b(t), u(0)=I,
    for t in (0,T] with steps of dt.
    a and b are Python functions of t.
    """
    dt = float(dt)          # avoid integer division
    Nt = int(round(T/dt))    # no of time intervals
    T = Nt*dt              # adjust T to fit time step dt
    u = zeros(Nt+1)         # array of u[n] values
    t = linspace(0, T, Nt+1) # time mesh

    u[0] = I                # assign initial condition
    for n in range(0, Nt):  # n=0,1,...,Nt-1
        u[n+1] = ((1 - dt*(1-theta)*a(t[n]))*u[n] + \
                  dt*(theta*b(t[n+1]) + (1-theta)*b(t[n]))) / \
                  (1 + dt*theta*a(t[n+1]))
    return u, t
```

This function is found in the file `decay_vc.py` (vc stands for "variable coefficients").

**Coding of variable coefficients.** The `solver` function shown above demands the arguments `a` and `b` to be Python functions of time `t`, say

```
def a(t):
    return a_0 if t < tp else k*a_0

def b(t):
    return 1
```

Here, `a(t)` has three parameters `a0`, `tp`, and `k`, which must be global variables. A better implementation is to represent `a` by a class where the parameters are attributes and a *special method* `__call__` evaluates  $a(t)$ :

```
class A:
    def __init__(self, a0=1, k=2):
        self.a0, self.k = a0, k

    def __call__(self, t):
        return self.a0 if t < self.tp else self.k*self.a0

a = A(a0=2, k=1) # a behaves as a function a(t)
```

For quick tests it is cumbersome to write a complete function or a class. The *lambda function* construction in Python is then convenient. For example,

```
a = lambda t: a_0 if t < tp else k*a_0
```

is equivalent to the `def a(t):` definition above. In general,

```
f = lambda arg1, arg2, ...: expressin
```

is equivalent to

```
def f(arg1, arg2, ...):  
    return expression
```

One can use lambda functions directly in calls. Say we want to solve  $u' = -u + 1$ ,  $u(0) = 2$ :

```
u, t = solver(2, lambda t: 1, lambda t: 1, T, dt, theta)
```

A lambda function can appear anywhere where a variable can appear.

## 6.4 Verifying a constant solution

An extremely useful partial verification method is to construct a test problem with a very simple solution, usually  $u = \text{const}$ . Especially the initial debugging of a program code can benefit greatly from such tests, because 1) all relevant numerical methods will exactly reproduce a constant solution, 2) many of the intermediate calculations are easy to control by hand for a constant  $u$ , and 3) even a constant  $u$  can uncover many bugs in an implementation.

The only constant solution for the problem  $u' = -au$  is  $u = 0$ , but too many bugs can escape from that trivial solution. It is much better to search for a problem where  $u = C = \text{const} \neq 0$ . Then  $u' = -a(t)u + b(t)$  is more appropriate: with  $u = C$  we can choose any  $a(t)$  and set  $b = a(t)C$  and  $I = C$ . An appropriate test function is

```
def test_constant_solution():  
    """  
    Test problem where u=u_const is the exact solution, to be  
    reproduced (to machine precision) by any relevant method.  
    """  
    def exact_solution(t):  
        return u_const  
  
    def a(t):  
        return 2.5*(1+t**3) # can be arbitrary  
  
    def b(t):  
        return a(t)*u_const  
  
    u_const = 2.15  
    theta = 0.4; I = u_const; dt = 4  
    Nt = 4 # enough with a few steps  
    u, t = solver(I=I, a=a, b=b, T=Nt*dt, dt=dt, theta=theta)  
    print u
```

```

u_e = exact_solution(t)
difference = abs(u_e - u).max() # max deviation
tol = 1E-14
assert difference < tol

```

An interesting question is what type of bugs that will make the computed  $u^n$  deviate from the exact solution  $C$ . Fortunately, the updating formula and the initial condition must be absolutely correct for the test to pass! Any attempt to make a wrong indexing in terms like  $\mathbf{a}(\mathbf{t}[\mathbf{n}])$  or any attempt to introduce an erroneous factor in the formula creates a solution that is different from  $C$ .

## 6.5 Verification via manufactured solutions

Following the idea of the previous section, we can choose any formula as the exact solution, insert the formula in the ODE problem and fit the data  $a(t)$ ,  $b(t)$ , and  $I$  to make the chosen formula fulfill the equation. This powerful technique for generating exact solutions is very useful for verification purposes and known as the *method of manufactured solutions*, often abbreviated MMS.

One common choice of solution is a linear function in the independent variable(s). The rationale behind such a simple variation is that almost any relevant numerical solution method for differential equation problems is able to reproduce the linear function exactly to machine precision (if  $u$  is about unity in size; precision is lost if  $u$  take on large values, see Exercise 8). The linear solution also makes some stronger demands to the numerical method and the implementation than the constant solution used in Section 6.4, at least in more complicated applications. However, the constant solution is often ideal for initial debugging before proceeding with a linear solution.

We choose a linear solution  $u(t) = ct + d$ . From the initial condition it follows that  $d = I$ . Inserting this  $u$  in the ODE results in

$$c = -a(t)u + b(t).$$

Any function  $u = ct + I$  is then a correct solution if we choose

$$b(t) = c + a(t)(ct + I).$$

With this  $b(t)$  there are no restrictions on  $a(t)$  and  $c$ .

Let prove that such a linear solution obeys the numerical schemes. To this end, we must check that  $u^n = ca(t_n)(ct_n + I)$  fulfills the discrete equations. For these calculations, and later calculations involving linear solutions inserted in finite difference schemes, it is convenient to compute the action of a difference operator on a linear function  $t$ :

$$[D_t^+ t]^n = \frac{t_{n+1} - t_n}{\Delta t} = 1, \quad (89)$$

$$[D_t^- t]^n = \frac{t_n - t_{n-1}}{\Delta t} = 1, \quad (90)$$

$$[D_t t]^n = \frac{t_{n+\frac{1}{2}} - t_{n-\frac{1}{2}}}{\Delta t} = \frac{(n + \frac{1}{2})\Delta t - (n - \frac{1}{2})\Delta t}{\Delta t} = 1. \quad (91)$$

Clearly, all three finite difference approximations to the derivative are exact for  $u(t) = t$  or its mesh function counterpart  $u^n = t_n$ .

The difference equation for the Forward Euler scheme

$$[D_t^+ u = -au + b]^n,$$

with  $a^n = a(t_n)$ ,  $b^n = c + a(t_n)(ct_n + I)$ , and  $u^n = ct_n + I$  then results in

$$c = -a(t_n)(ct_n + I) + c + a(t_n)(ct_n + I) = c$$

which is always fulfilled. Similar calculations can be done for the Backward Euler and Crank-Nicolson schemes, or the  $\theta$ -rule for that matter. In all cases,  $u^n = ct_n + I$  is an exact solution of the discrete equations. That is why we should expect that  $u^n - u_e(t_n) = 0$  mathematically and  $|u^n - u_e(t_n)|$  less than a small number about the machine precision for  $n = 0, \dots, N_t$ .

The following function offers an implementation of this verification test based on a linear exact solution:

```
def test_linear_solution():
    """
    Test problem where u=c*t+I is the exact solution, to be
    reproduced (to machine precision) by any relevant method.
    """
    def exact_solution(t):
        return c*t + I

    def a(t):
        return t**0.5 # can be arbitrary

    def b(t):
        return c + a(t)*exact_solution(t)

    theta = 0.4; I = 0.1; dt = 0.1; c = -0.5
    T = 4
    Nt = int(T/dt) # no of steps
    u, t = solver(I=I, a=a, b=b, T=Nt*dt, dt=dt, theta=theta)
    u_e = exact_solution(t)
    difference = abs(u_e - u).max() # max deviation
    print difference
    tol = 1E-14 # depends on c!
    assert difference < tol
```

Any error in the updating formula makes this test fail!

Choosing more complicated formulas as the exact solution, say  $\cos(t)$ , will not make the numerical and exact solution coincide to machine precision, because finite differencing of  $\cos(t)$  does not exactly yield the exact derivative  $-\sin(t)$ . In such cases, the verification procedure must be based on measuring the convergence rates as exemplified in Section 6.6. Convergence rates can be computed as long as one has an exact solution of a problem that the solver can be tested on, but this can always be obtained by the method of manufactured solutions.

## 6.6 Computing convergence rates

We expect that the error  $E$  in the numerical solution is reduced if the mesh size  $\Delta t$  is decreased. More specifically, many numerical methods obey a power-law relation between  $E$  and  $\Delta t$ :

$$E = C\Delta t^r, \quad (92)$$

where  $C$  and  $r$  are (usually unknown) constants independent of  $\Delta t$ . The formula (92) is viewed as an asymptotic model valid for sufficiently small  $\Delta t$ . How small is normally hard to estimate without doing numerical estimations of  $r$ .

The parameter  $r$  is known as the *convergence rate*. For example, if the convergence rate is 2, halving  $\Delta t$  reduces the error by a factor of 4. Diminishing  $\Delta t$  then has a greater impact on the error compared with methods that have  $r = 1$ . For a given value of  $r$ , we refer to the method as of  $r$ -th order. First- and second-order methods are most common in scientific computing.

**Estimating  $r$ .** There are two alternative ways of estimating  $C$  and  $r$  based on a set of  $m$  simulations with corresponding pairs  $(\Delta t_i, E_i)$ ,  $i = 0, \dots, m-1$ , and  $\Delta t_i < \Delta t_{i-1}$  (i.e., decreasing cell size).

1. Take the logarithm of (92),  $\ln E = r \ln \Delta t + \ln C$ , and fit a straight line to the data points  $(\Delta t_i, E_i)$ ,  $i = 0, \dots, m-1$ .
2. Consider two consecutive experiments,  $(\Delta t_i, E_i)$  and  $(\Delta t_{i-1}, E_{i-1})$ . Dividing the equation  $E_{i-1} = C\Delta t_{i-1}^r$  by  $E_i = C\Delta t_i^r$  and solving for  $r$  yields

$$r_{i-1} = \frac{\ln(E_{i-1}/E_i)}{\ln(\Delta t_{i-1}/\Delta t_i)} \quad (93)$$

for  $i = 1, \dots, m-1$ .

The disadvantage of method 1 is that (92) might not be valid for the coarsest meshes (largest  $\Delta t$  values). Fitting a line to all the data points is then misleading. Method 2 computes convergence rates for pairs of experiments and allows us to see if the sequence  $r_i$  converges to some value as  $i \rightarrow m-2$ . The final  $r_{m-2}$  can then be taken as the convergence rate. If the coarsest meshes have a differing rate, the corresponding time steps are probably too large for (92) to be valid. That is, those time steps lie outside the asymptotic range of  $\Delta t$  values where the error behaves like (92).

**Implementation.** Suppose we have some function `error(dt, ...)` that can compute the numerical error in our mathematical model. To compute  $r_0, r_1, \dots, r_{m-2}$  from (92) we just wrap a loop over  $\Delta t$  values (`dt_values`) around the `error` function:

```

E_values = [error(dt, ...) for dt in dt_values]

# Compute pairwise convergence rates
m = len(dt_values)
r = [log(E_values[i-1]/E_values[i])/
      log(dt_values[i-1]/dt_values[i])
      for i in range(1, m, 1)]

# Strip off to 2 decimals
r = [round(r_, 2) for r_ in r]

```

We can run the convergence rate estimate computations for the  $\theta$ -rule discretization of  $u' = -au$ , using  $\Delta t = 0.5, -0.25, 0.1, 0.05, 0.025, 0.01$ :

```

FE: 1.33 1.15 1.07 1.03 1.02
BE: 0.98 0.99 0.99 1.00 1.00
CN: 2.14 2.07 2.03 2.01 2.01

```

The Forward and Backward Euler methods seem to have an  $r$  value which stabilizes at 1, while the Crank-Nicolson seems to be a second-order method with  $r = 2$ . These results are in very good agreement with various theoretical considerations for  $r$ .

#### Why convergence rates are important.

The strong practical application of computing convergence rates is for verification: wrong convergence rates point to errors in the code, and correct convergence rates brings evidence that the implementation is correct. Experience shows that bugs in the code easily destroy the expected convergence rate.

## 6.7 Extension to systems of ODEs

Many ODE models involves more than one unknown function and more than one equation. Here is an example of two unknown functions  $u(t)$  and  $v(t)$ :

$$u' = au + bv, \quad (94)$$

$$v' = cu + dv, \quad (95)$$

for constants  $a, b, c, d$ . Applying the Forward Euler method to each equation results in simple updating formula

$$u^{n+1} = u^n + \Delta t(au^n + bv^n), \quad (96)$$

$$v^{n+1} = v^n + \Delta t(cu^n + dv^n). \quad (97)$$

On the other hand, the Crank-Nicolson or Backward Euler schemes result in a  $2 \times 2$  linear system for the new unknowns. The latter schemes gives

$$u^{n+1} = u^n + \Delta t(au^{n+1} + bv^{n+1}), \quad (98)$$

$$v^{n+1} = v^n + \Delta t(cu^{n+1} + dv^{n+1}). \quad (99)$$

Collecting  $u^{n+1}$  as well as  $v^{n+1}$  on the left-hand side results in

$$(1 - \Delta ta)u^{n+1} + bv^{n+1} = u^n, \quad (100)$$

$$cu^{n+1} + (1 - \Delta td)v^{n+1} = v^n, \quad (101)$$

which is a system of two coupled, linear, algebraic equations in two unknowns.

## 7 General first-order ODEs

We now turn the attention to general, nonlinear ODEs and systems of such ODEs. Our focus is on numerical methods that can be readily reused for time-discretization PDEs, and diffusion PDEs in particular. The methods are just briefly listed, and we refer to the rich literature for more detailed descriptions and analysis - the books [7, 1, 2, 3] are all excellent resources on numerical methods for ODEs. We also demonstrate the Odespy Python interface to a range of different software for general first-order ODE systems.

### 7.1 Generic form of first-order ODEs

ODEs are commonly written in the generic form

$$u' = f(u, t), \quad u(0) = I, \quad (102)$$

where  $f(u, t)$  is some prescribed function. As an example, our most general exponential decay model (80) has  $f(u, t) = -a(t)u(t) + b(t)$ .

The unknown  $u$  in (102) may either be a scalar function of time  $t$ , or a vector valued function of  $t$  in case of a *system of ODEs* with  $m$  unknown components:

$$u(t) = (u^{(0)}(t), u^{(1)}(t), \dots, u^{(m-1)}(t)).$$

In that case, the right-hand side is vector-valued function with  $m$  components,

$$\begin{aligned} f(u, t) = & (f^{(0)}(u^{(0)}(t), \dots, u^{(m-1)}(t)), \\ & f^{(1)}(u^{(0)}(t), \dots, u^{(m-1)}(t)), \\ & \vdots, \\ & f^{(m-1)}(u^{(0)}(t), \dots, u^{(m-1)}(t))). \end{aligned}$$

Actually, any system of ODEs can be written in the form (102), but higher-order ODEs then need auxiliary unknown functions to enable conversion to a first-order system.

Next we list some well-known methods for  $u' = f(u, t)$ , valid both for a single ODE (scalar  $u$ ) and systems of ODEs (vector  $u$ ). The choice of methods is inspired by the kind of schemes that are popular also for time discretization of partial differential equations.

## 7.2 The $\theta$ -rule

The  $\theta$ -rule scheme applied to  $u' = f(u, t)$  becomes

$$\frac{u^{n+1} - u^n}{\Delta t} = \theta f(u^{n+1}, t_{n+1}) + (1 - \theta)f(u^n, t_n). \quad (103)$$

Bringing the unknown  $u^{n+1}$  to the left-hand side and the known terms on the right-hand side gives

$$u^{n+1} - \Delta t \theta f(u^{n+1}, t_{n+1}) = u^n + \Delta t (1 - \theta) f(u^n, t_n). \quad (104)$$

For a general  $f$  (not linear in  $u$ ), this equation is *nonlinear* in the unknown  $u^{n+1}$  unless  $\theta = 0$ . For a scalar ODE ( $m = 1$ ), we have to solve a single nonlinear algebraic equation for  $u^{n+1}$ , while for a system of ODEs, we get a system of coupled, nonlinear algebraic equations. Newton's method is a popular solution approach in both cases. Note that with the Forward Euler scheme ( $\theta = 0$ ) we do not have to deal with nonlinear equations, because in that case we have an explicit updating formula for  $u^{n+1}$ . This is known as an *explicit* scheme. With  $\theta \neq 1$  we have to solve (systems of) algebraic equations, and the scheme is said to be *implicit*.

## 7.3 An implicit 2-step backward scheme

The implicit backward method with 2 steps applies a three-level backward difference as approximation to  $u'(t)$ ,

$$u'(t_{n+1}) \approx \frac{3u^{n+1} - 4u^n + u^{n-1}}{2\Delta t},$$

which is an approximation of order  $\Delta t^2$  to the first derivative. The resulting scheme for  $u' = f(u, t)$  reads

$$u^{n+1} = \frac{4}{3}u^n - \frac{1}{3}u^{n-1} + \frac{2}{3}\Delta t f(u^{n+1}, t_{n+1}). \quad (105)$$

Higher-order versions of the scheme (105) can be constructed by including more time levels. These schemes are known as the Backward Differentiation Formulas (BDF), and the particular version (105) is often referred to as BDF2.

Note that the scheme (105) is implicit and requires solution of nonlinear equations when  $f$  is nonlinear in  $u$ . The standard 1st-order Backward Euler method or the Crank-Nicolson scheme can be used for the first step.



## 7.4 Leapfrog schemes

**The ordinary Leapfrog scheme.** The derivative of  $u$  at some point  $t_n$  can be approximated by a central difference over two time steps,

$$u'(t_n) \approx \frac{u^{n+1} - u^{n-1}}{2\Delta t} = [D_{2t}u]^n \quad (106)$$

which is an approximation of second order in  $\Delta t$ . The scheme can then be written as

$$[D_{2t}u = f(u, t)]^n,$$

in operator notation. Solving for  $u^{n+1}$  gives

$$u^{n+1} = u^{n-1} + \Delta t f(u^n, t_n). \quad (107)$$

Observe that (107) is an explicit scheme, and that a nonlinear  $f$  (in  $u$ ) is trivial to handle since it only involves the known  $u^n$  value. Some other scheme must be used as starter to compute  $u^1$ , preferably the Forward Euler scheme since it is also explicit.

**The filtered Leapfrog scheme.** Unfortunately, the Leapfrog scheme (107) will develop growing oscillations with time (see Problem 13). A remedy for such undesired oscillations is to introduce a *filtering technique*. First, a standard Leapfrog step is taken, according to (107), and then the previous  $u^n$  value is adjusted according to

$$u^n \leftarrow u^n + \gamma(u^{n-1} - 2u^n + u^{n+1}). \quad (108)$$

The  $\gamma$ -terms will effectively damp oscillations in the solution, especially those with short wavelength (like point-to-point oscillations). A common choice of  $\gamma$  is 0.6 (a value used in the famous NCAR Climate Model).

## 7.5 The 2nd-order Runge-Kutta method

The two-step scheme

$$u^* = u^n + \Delta t f(u^n, t_n), \quad (109)$$

$$u^{n+1} = u^n + \Delta t \frac{1}{2} (f(u^n, t_n) + f(u^*, t_{n+1})), \quad (110)$$

essentially applies a Crank-Nicolson method (110) to the ODE, but replaces the term  $f(u^{n+1}, t_{n+1})$  by a prediction  $f(u^*, t_{n+1})$  based on a Forward Euler step (109). The scheme (109)-(110) is known as Huen's method, but is also a 2nd-order Runge-Kutta method. The scheme is explicit, and the error is expected to behave as  $\Delta t^2$ .

## 7.6 A 2nd-order Taylor-series method

One way to compute  $u^{n+1}$  given  $u^n$  is to use a Taylor polynomial. We may write up a polynomial of 2nd degree:

$$u^{n+1} = u^n + u'(t_n)\Delta t + \frac{1}{2}u''(t_n)\Delta t^2.$$

From the equation  $u' = f(u, t)$  it follows that the derivatives of  $u$  can be expressed in terms of  $f$  and its derivatives:

$$\begin{aligned} u'(t_n) &= f(u^n, t_n), \\ u''(t_n) &= \frac{\partial f}{\partial u}(u^n, t_n)u'(t_n) + \frac{\partial f}{\partial t} \\ &= f(u^n, t_n)\frac{\partial f}{\partial u}(u^n, t_n) + \frac{\partial f}{\partial t}, \end{aligned}$$

resulting in the scheme

$$u^{n+1} = u^n + f(u^n, t_n)\Delta t + \frac{1}{2}\left(f(u^n, t_n)\frac{\partial f}{\partial u}(u^n, t_n) + \frac{\partial f}{\partial t}\right)\Delta t^2. \quad (111)$$

More terms in the series could be included in the Taylor polynomial to obtain methods of higher order than 2.

## 7.7 The 2nd- and 3rd-order Adams-Bashforth schemes

The following method is known as the 2nd-order Adams-Bashforth scheme:

$$u^{n+1} = u^n + \frac{1}{2}\Delta t \left(3f(u^n, t_n) - f(u^{n-1}, t_{n-1})\right). \quad (112)$$

The scheme is explicit and requires another one-step scheme to compute  $u^1$  (the Forward Euler scheme or Heun's method, for instance). As the name implies, the scheme is of order  $\Delta t^2$ .

Another explicit scheme, involving four time levels, is the 3rd-order Adams-Bashforth scheme

$$u^{n+1} = u^n + \frac{1}{12}\Delta t \left(23f(u^n, t_n) - 16f(u^{n-1}, t_{n-1}) + 5f(u^{n-2}, t_{n-2})\right). \quad (113)$$

The numerical error is of order  $\Delta t^3$ , and the scheme needs some method for computing  $u^1$  and  $u^2$ .

More general, higher-order Adams-Bashforth schemes (also called *explicit Adams methods*) compute  $u^{n+1}$  as a linear combination of  $f$  at  $k$  previous time steps:

$$u^{n+1} = u^n + \sum_{j=0}^k \beta_j f(u^{n-j}, t_{n-j}),$$

where  $\beta_j$  are known coefficients.

## 7.8 The 4th-order Runge-Kutta method

The perhaps most widely used method to solve ODEs is the 4th-order Runge-Kutta method, often called RK4. Its derivation is a nice illustration of common numerical approximation strategies, so let us go through the steps in detail.

The starting point is to integrate the ODE  $u' = f(u, t)$  from  $t_n$  to  $t_{n+1}$ :

$$u(t_{n+1}) - u(t_n) = \int_{t_n}^{t_{n+1}} f(u(t), t) dt.$$

We want to compute  $u(t_{n+1})$  and regard  $u(t_n)$  as known. The task is to find good approximations for the integral, since the integrand involves the unknown  $u$  between  $t_n$  and  $t_{n+1}$ .

The integral can be approximated by the famous [Simpson's rule](#):

$$\int_{t_n}^{t_{n+1}} f(u(t), t) dt \approx \frac{\Delta t}{6} \left( f^n + 4f^{n+\frac{1}{2}} + f^{n+1} \right).$$

The problem now is that we do not know  $f^{n+\frac{1}{2}} = f(u^{n+\frac{1}{2}}, t_{n+\frac{1}{2}})$  and  $f^{n+1} = f(u^{n+1}, t_{n+1})$  as we know only  $u^n$  and hence  $f^n$ . The idea is to use various approximations for  $f^{n+\frac{1}{2}}$  and  $f^{n+1}$  based on using well-known schemes for the ODE in the intervals  $[t_n, t_{n+\frac{1}{2}}]$  and  $[t_{n+\frac{1}{2}}, t_{n+1}]$ . We split the integral approximation into four terms:

$$\int_{t_n}^{t_{n+1}} f(u(t), t) dt \approx \frac{\Delta t}{6} \left( f^n + 2\hat{f}^{n+\frac{1}{2}} + 2\tilde{f}^{n+\frac{1}{2}} + \bar{f}^{n+1} \right),$$

where  $\hat{f}^{n+\frac{1}{2}}$ ,  $\tilde{f}^{n+\frac{1}{2}}$ , and  $\bar{f}^{n+1}$  are approximations to  $f^{n+\frac{1}{2}}$  and  $f^{n+1}$  that can be based on already computed quantities. For  $\hat{f}^{n+\frac{1}{2}}$  we can apply an approximation to  $u^{n+\frac{1}{2}}$  using the Forward Euler method with step  $\frac{1}{2}\Delta t$ :

$$\hat{f}^{n+\frac{1}{2}} = f\left(u^n + \frac{1}{2}\Delta t f^n, t_{n+\frac{1}{2}}\right) \quad (114)$$

Since this gives us a prediction of  $f^{n+\frac{1}{2}}$ , we can for  $\tilde{f}^{n+\frac{1}{2}}$  try a Backward Euler method to approximate  $u^{n+\frac{1}{2}}$ :

$$\tilde{f}^{n+\frac{1}{2}} = f\left(u^n + \frac{1}{2}\Delta t \hat{f}^{n+\frac{1}{2}}, t_{n+\frac{1}{2}}\right). \quad (115)$$

With  $\tilde{f}^{n+\frac{1}{2}}$  as a hopefully good approximation to  $f^{n+\frac{1}{2}}$ , we can for the final term  $\bar{f}^{n+1}$  use a Crank-Nicolson method to approximate  $u^{n+1}$ :

$$\bar{f}^{n+1} = f\left(u^n + \Delta t \tilde{f}^{n+\frac{1}{2}}, t_{n+1}\right). \quad (116)$$

We have now used the Forward and Backward Euler methods as well as the Crank-Nicolson method in the context of Simpson's rule. The hope is that the

combination of these methods yields an overall time-stepping scheme from  $t_n$  to  $t_{n+1}$  that is much more accurate than the  $\mathcal{O}(\Delta t)$  and  $\mathcal{O}(\Delta t^2)$  of the individual steps. This is indeed true: the overall accuracy is  $\mathcal{O}(\Delta t^4)$ !

To summarize, the 4th-order Runge-Kutta method becomes

$$u^{n+1} = u^n + \frac{\Delta t}{6} \left( f^n + 2\hat{f}^{n+\frac{1}{2}} + 2\tilde{f}^{n+\frac{1}{2}} + \bar{f}^{n+1} \right), \quad (117)$$

where the quantities on the right-hand side are computed from (114)-(116). Note that the scheme is fully explicit so there is never any need to solve linear or nonlinear algebraic equations. However, the stability is conditional and depends on  $f$ . There is a whole range of *implicit* Runge-Kutta methods that are unconditionally stable, but require solution of algebraic equations involving  $f$  at each time step.

The simplest way to explore more sophisticated methods for ODEs is to apply one of the many high-quality software packages that exist, as the next section explains.

## 7.9 The Odespy software

A wide range of the methods and software exist for solving (102). Many of methods are accessible through a unified Python interface offered by the [Odespy](#) [5] package. Odespy features simple Python implementations of the most fundamental schemes as well as Python interfaces to several famous packages for solving ODEs: [ODEPACK](#), [Vode](#), [rk4.f](#), [rkf45.f](#), [Radau5](#), as well as the ODE solvers in [SciPy](#), [SymPy](#), and [odelab](#).

The usage of Odespy follows this setup for the ODE  $u' = -au$ ,  $u(0) = I$ ,  $t \in (0, T]$ , here solved by the famous 4th-order Runge-Kutta method, using  $\Delta t = 1$  and  $N_t = 6$  steps:

```
def f(u, t):
    return -a*u

import odespy
import numpy as np

I = 1; a = 0.5; Nt = 6; dt = 1
solver = odespy.RK4(f)
solver.set_initial_condition(I)
t_mesh = np.linspace(0, Nt*dt, Nt+1)
u, t = solver.solve(t_mesh)
```

The previously listed methods for ODEs are all accessible in Odespy:

- the  $\theta$ -rule: `ThetaRule`
- special cases of the  $\theta$ -rule: `ForwardEuler`, `BackwardEuler`, `CrankNicolson`
- the 2nd- and 4th-order Runge-Kutta methods: `RK2` and `RK4`
- The BDF methods and the Adam-Bashforth methods: `Vode`, `Lsode`, `Lsoda`, `lsoda_scipy`
- The Leapfrog scheme: `Leapfrog` and `LeapfrogFiltered`

## 7.10 Example: Runge-Kutta methods

Since all solvers have the same interface in Odespy, modulo different set of parameters to the solvers' constructors, one can easily make a list of solver objects and run a loop for comparing (a lot of) solvers. The code below, found in complete form in `decay_odespy.py`, compares the famous Runge-Kutta methods of orders 2, 3, and 4 with the exact solution of the decay equation  $u' = -au$ . Since we have quite long time steps, we have included the only relevant  $\theta$ -rule for large time steps, the Backward Euler scheme ( $\theta = 1$ ), as well. Figure 16 shows the results.

```
import numpy as np
import matplotlib.pyplot as plt
import sys

def f(u, t):
    return -a*u

I = 1; a = 2; T = 6
dt = float(sys.argv[1]) if len(sys.argv) >= 2 else 0.75
Nt = int(round(T/dt))
t = np.linspace(0, Nt*dt, Nt+1)

solvers = [odespy.RK2(f),
            odespy.RK3(f),
            odespy.RK4(f),]

# BackwardEuler must use Newton solver to converge
# (Picard is default and leads to divergence)
solvers.append(
    odespy.BackwardEuler(f, nonlinear_solver='Newton'))
# Or tell BackwardEuler that it is a linear problem
solvers[-1] = odespy.BackwardEuler(f, f_is_linear=True,
                                   jac=lambda u, t: -a)

legends = []
for solver in solvers:
    solver.set_initial_condition(I)
    u, t = solver.solve(t)

    plt.plot(t, u)
    plt.hold('on')
    legends.append(solver.__class__.__name__)

# Compare with exact solution plotted on a very fine mesh
t_fine = np.linspace(0, T, 10001)
u_e = I*np.exp(-a*t_fine)
plt.plot(t_fine, u_e, '-') # avoid markers by specifying line type
legends.append('exact')

plt.legend(legends)
plt.title('Time step: %g' % dt)
plt.show()
```

With the `odespy.BackwardEuler` method we either have to tell that the problem is linear and provide the Jacobian of  $f(u, t)$ , i.e.,  $\partial f / \partial u$ , as the `jac` argument, or we have to assume that  $f$  is nonlinear, but then specify Newton's method as solver for the nonlinear equations (since the equations are linear, Newton's

method will converge in one iteration). By default, `odespy.BackwardEuler` assumes a nonlinear problem to be solved by Picard iteration, but that leads to divergence in the present problem.

#### Visualization tip.

We use Matplotlib for plotting here, but importing `scitools.std` as `plt` instead also works. Plain use of Matplotlib as done here results in curves with different colors, which may be hard to distinguish on black-and-white paper. Using `scitools.std`, curves are automatically given colors *and* markers, thus making curves easy to distinguish on screen with colors and on black-and-white paper. The automatic adding of markers is normally a bad idea for a very fine mesh since all the markers get cluttered, but `scitools.std` limits the number of markers in such cases. For the exact solution we use a very fine mesh, but in the code above we specify the line type as a solid line (-), which means no markers and just a color to be automatically determined by the backend used for plotting (Matplotlib by default, but `scitools.std` gives the opportunity to use other backends to produce the plot, e.g., Gnuplot or Grace).

Also note that the legends are based on the class names of the solvers, and in Python the name of a the class type (as a string) of an object `obj` is obtained by `obj.__class__.__name__`.

The runs in Figure 16 and other experiments reveal that the 2nd-order Runge-Kutta method (RK2) is unstable for  $\Delta t > 1$  and decays slower than the Backward Euler scheme for large and moderate  $\Delta t$  (see Exercise 12 for an analysis). However, for fine  $\Delta t = 0.25$  the 2nd-order Runge-Kutta method approaches the exact solution faster than the Backward Euler scheme. That is, the latter scheme does a better job for larger  $\Delta t$ , while the higher order scheme is superior for smaller  $\Delta t$ . This is a typical trend also for most schemes for ordinary and partial differential equations.

The 3rd-order Runge-Kutta method (RK3) has also artifacts in form of oscillatory behavior for the larger  $\Delta t$  values, much like that of the Crank-Nicolson scheme. For finer  $\Delta t$ , the 3rd-order Runge-Kutta method converges quickly to the exact solution.

The 4th-order Runge-Kutta method (RK4) is slightly inferior to the Backward Euler scheme on the coarsest mesh, but is then clearly superior to all the other schemes. It is definitely the method of choice for all the tested schemes.

**Remark about using the  $\theta$ -rule in Odespy.** The Odespy package assumes that the ODE is written as  $u' = f(u, t)$  with an  $f$  that is possibly nonlinear in  $u$ . The  $\theta$ -rule for  $u' = f(u, t)$  leads to

$$u^{n+1} = u^n + \Delta t \left( \theta f(u^{n+1}, t_{n+1}) + (1 - \theta) f(u^n, t_n) \right),$$

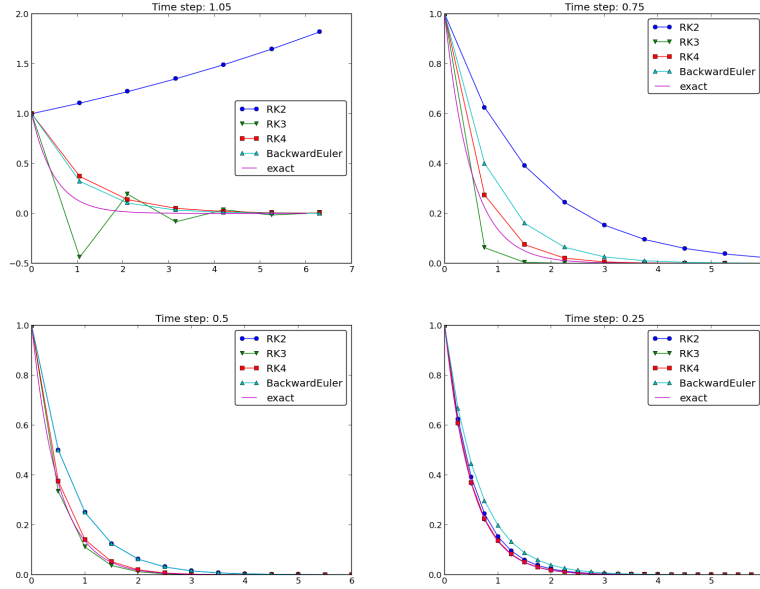


Figure 16: Behavior of different schemes for the decay equation.

which is a *nonlinear equation* in  $u^{n+1}$ . Odespy's implementation of the  $\theta$ -rule (**ThetaRule**) and the specialized Backward Euler (**BackwardEuler**) and Crank-Nicolson (**CrankNicolson**) schemes must invoke iterative methods for solving the nonlinear equation in  $u^{n+1}$ . This is done even when  $f$  is linear in  $u$ , as in the model problem  $u' = -au$ , where we can easily solve for  $u^{n+1}$  by hand. Therefore, we need to specify use of Newton's method to solve the equations. (Odespy allows other methods than Newton's to be used, for instance Picard iteration, but that method is not suitable. The reason is that it applies the Forward Euler scheme to generate a start value for the iterations. Forward Euler may give very wrong solutions for large  $\Delta t$  values. Newton's method, on the other hand, is insensitive to the start value in *linear problems*.)

### 7.11 Example: Adaptive Runge-Kutta methods

Odespy offers solution methods that can adapt the size of  $\Delta t$  with time to match a desired accuracy in the solution. Intuitively, small time steps will be chosen in areas where the solution is changing rapidly, while larger time steps can be used where the solution is slowly varying. Some kind of *error estimator* is used to adjust the next time step at each time level.

A very popular adaptive method for solving ODEs is the Dormand-Prince Runge-Kutta method of order 4 and 5. The 5th-order method is used as a reference solution and the difference between the 4th- and 5th-order methods is

used as an indicator of the error in the numerical solution. The Dormand-Prince method is the default choice in MATLAB's widely used `ode45` routine.

We can easily set up Odespy to use the Dormand-Prince method and see how it selects the optimal time steps. To this end, we request only one time step from  $t = 0$  to  $t = T$  and ask the method to compute the necessary non-uniform time mesh to meet a certain error tolerance. The code goes like

```
import odespy
import numpy as np
import decay_mod
import sys
#import matplotlib.pyplot as plt
import scitools.std as plt

def f(u, t):
    return -a*u

def exact_solution(t):
    return I*np.exp(-a*t)

I = 1; a = 2; T = 5
tol = float(sys.argv[1])
solver = odespy.DormandPrince(f, atol=tol, rtol=0.1*tol)

Nt = 1 # just one step - let the scheme find its intermediate points
t_mesh = np.linspace(0, T, Nt+1)
t_fine = np.linspace(0, T, 10001)

solver.set_initial_condition(I)
u, t = solver.solve(t_mesh)

# u and t will only consist of [I, u^Nt] and [0,T]
# solver.u_all and solver.t_all contains all computed points
plt.plot(solver.t_all, solver.u_all, 'ko')
plt.hold('on')
plt.plot(t_fine, exact_solution(t_fine), 'b-')
plt.legend(['tol=%0E' % tol, 'exact'])
plt.savefig('tmp_odespy_adaptive.png')
plt.show()
```

Running four cases with tolerances  $10^{-1}$ ,  $10^{-3}$ ,  $10^{-5}$ , and  $10^{-7}$ , gives the results in Figure 17. Intuitively, one would expect denser points in the beginning of the decay and larger time steps when the solution flattens out.



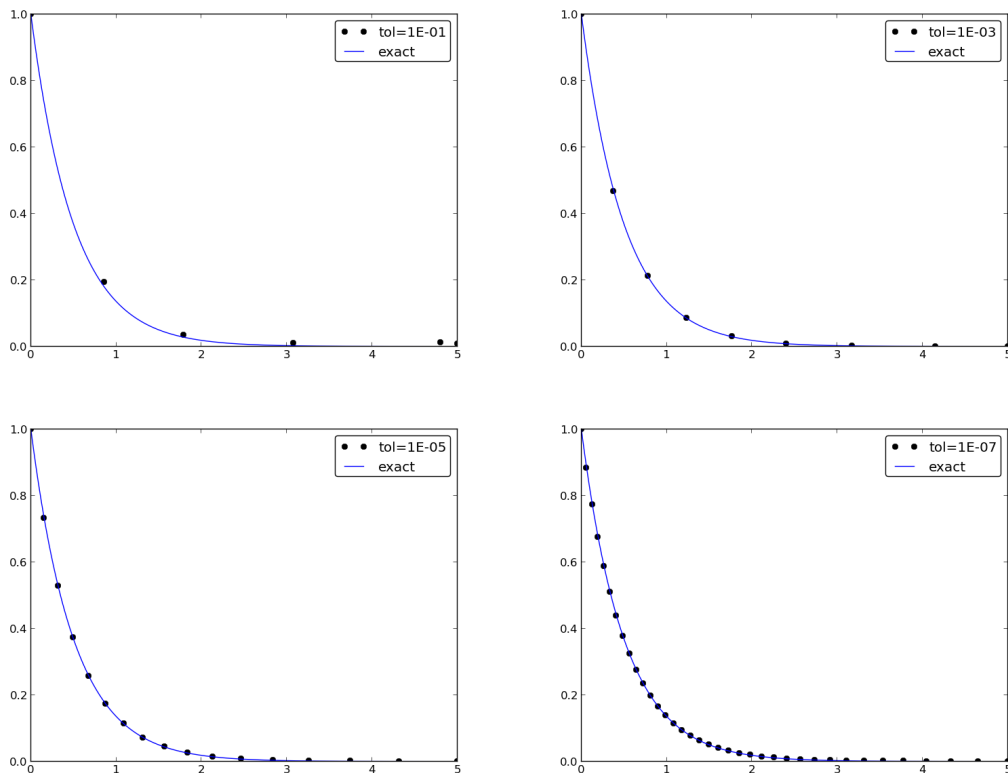


Figure 17: Choice of adaptive time mesh by the Dormand-Prince method for different tolerances.

## 8 Exercises

### Exercise 8: Experiment with precision in tests and the size of $u$

It is claimed in Section 6.5 that most numerical methods will reproduce a linear exact solution to machine precision. Test this assertion using the test function `test_linear_solution` in the `decay_vc.py` program. Vary the parameter `c` from very small, via `c=1` to many larger values, and print out the maximum difference between the numerical solution and the exact solution. What is the relevant value of the tolerance in the float comparison in each case? Filename: `test_precision.py`.

### Exercise 9: Implement the 2-step backward scheme

Implement the 2-step backward method (105) for the model  $u'(t) = -a(t)u(t) + b(t)$ ,  $u(0) = I$ . Allow the first step to be computed by either the Backward Euler scheme or the Crank-Nicolson scheme. Verify the implementation by choosing  $a(t)$  and  $b(t)$  such that the exact solution is linear in  $t$  (see Section 6.5). Show mathematically that a linear solution is indeed a solution of the discrete equations.

Compute convergence rates (see Section 6.6) in a test case  $a = \text{const}$  and  $b = 0$ , where we easily have an exact solution, and determine if the choice of a first-order scheme (Backward Euler) for the first step has any impact on the overall accuracy of this scheme. The expected error goes like  $\mathcal{O}(\Delta t^2)$ . Filename: `decay_backward2step.py`.

### Exercise 10: Implement the 2nd-order Adams-Bashforth scheme

Implement the 2nd-order Adams-Bashforth method (112) for the decay problem  $u' = -a(t)u + b(t)$ ,  $u(0) = I$ ,  $t \in (0, T]$ . Use the Forward Euler method for the first step such that the overall scheme is explicit. Verify the implementation using an exact solution that is linear in time. Analyze the scheme by searching for solutions  $u^n = A^n$  when  $a = \text{const}$  and  $b = 0$ . Compare this second-order scheme to the Crank-Nicolson scheme. Filename: `decay_AdamsBashforth2.py`.

### Exercise 11: Implement the 3rd-order Adams-Bashforth scheme

Implement the 3rd-order Adams-Bashforth method (113) for the decay problem  $u' = -a(t)u + b(t)$ ,  $u(0) = I$ ,  $t \in (0, T]$ . Since the scheme is explicit, allow it to be started by two steps with the Forward Euler method. Investigate experimentally the case where  $b = 0$  and  $a$  is a constant: Can we have oscillatory solutions for large  $\Delta t$ ? Filename: `decay_AdamsBashforth3.py`.

### Exercise 12: Analyze explicit 2nd-order methods

Show that the schemes (110) and (111) are identical in the case  $f(u, t) = -a$ , where  $a > 0$  is a constant. Assume that the numerical solution reads  $u^n = A^n$  for some unknown amplification factor  $A$  to be determined. Find  $A$  and derive stability criteria. Can the scheme produce oscillatory solutions of  $u' = -au$ ? Plot the numerical and exact amplification factor. Filename: `decay_RK2_Taylor2.py`.

### Problem 13: Implement and investigate the Leapfrog scheme

A Leapfrog scheme for the ODE  $u'(t) = -a(t)u(t) + b(t)$  is defined by

$$[D_{2t}u = -au + b]^n.$$

A separate method is needed to compute  $u^1$ . The Forward Euler scheme is a possible candidate.

**a)** Implement the Leapfrog scheme for the model equation. Plot the solution in the case  $a = 1$ ,  $b = 0$ ,  $I = 1$ ,  $\Delta t = 0.01$ ,  $t \in [0, 4]$ . Compare with the exact solution  $u_e(t) = e^{-t}$ .

**b)** Show mathematically that a linear solution in  $t$  fulfills the Forward Euler scheme for the first step and the Leapfrog scheme for the subsequent steps. Use this linear solution to verify the implementation, and automate the verification through a test function.

**Hint.** It can be wise to automate the calculations such that it is easy to redo the calculations for other types of solutions. Here is a possible `sympy` function that takes a symbolic expression `u` (implemented as a Python function of `t`), fits the `b` term, and checks if `u` fulfills the discrete equations:

```
import sympy as sp

def analyze(u):
    t, dt, a = sp.symbols('t dt a')

    print 'Analyzing u_e(t)=%s' % u(t)
    print 'u(0)=%s' % u(t).subs(t, 0)

    # Fit source term to the given u(t)
    b = sp.diff(u(t), t) + a*u(t)
    b = sp.simplify(b)
    print 'Source term b:', b

    # Residual in discrete equations; Forward Euler step
    R_step1 = (u(t+dt) - u(t))/dt + a*u(t) - b
    R_step1 = sp.simplify(R_step1)
    print 'Residual Forward Euler step:', R_step1

    # Residual in discrete equations; Leapfrog steps
    R = (u(t+dt) - u(t-dt))/(2*dt) + a*u(t) - b
    R = sp.simplify(R)
    print 'Residual Leapfrog steps:', R

def u_e(t):
    return c*t + I

analyze(u_e)
# or short form: analyze(lambda t: c*t + I)
```

**c)** Show that a second-order polynomial in  $t$  cannot be a solution of the discrete equations. However, if a Crank-Nicolson scheme is used for the first step, a second-order polynomial solves the equations exactly.

d) Create a manufactured solution  $u(t) = \sin(t)$  for the ODE  $u' = -au + b$ . Compute the convergence rate of the Leapfrog scheme using this manufactured solution. The expected convergence rate of the Leapfrog scheme is  $\mathcal{O}(\Delta t^2)$ . Does the use of a 1st-order method for the first step impact the convergence rate?

e) Set up a set of experiments to demonstrate that the Leapfrog scheme (107) is associated with numerical artifacts (instabilities). Document the main results from this investigation.

f) Analyze and explain the instabilities of the Leapfrog scheme (107):

1. Choose  $a = \text{const}$  and  $b = 0$ . Assume that an exact solution of the discrete equations has the form  $u^n = A^n$ , where  $A$  is an amplification factor to be determined. Derive an equation for  $A$  by inserting  $u^n = A^n$  in the Leapfrog scheme.
2. Compute  $A$  either by hand and/or with the aid of `sympy`. The polynomial for  $A$  has two roots,  $A_1$  and  $A_2$ . Let  $u^n$  be a linear combination  $u^n = C_1 A_1^n + C_2 A_2^n$ .
3. Show that one of the roots is the explanation of the instability.
4. Compare  $A$  with the exact expression, using a Taylor series approximation.
5. How can  $C_1$  and  $C_2$  be determined?

g) Since the original Leapfrog scheme is unconditionally unstable as time grows, it demands some stabilization. This can be done by filtering, where we first find  $u^{n+1}$  from the original Leapfrog scheme and then replace  $u^n$  by  $u^n + \gamma(u^{n-1} - 2u^n + u^{n+1})$ , where  $\gamma$  can be taken as 0.6. Implement the filtered Leapfrog scheme and check that it can handle tests where the original Leapfrog scheme is unstable.

Filenames: `decay_leapfrog.py`, `decay_leapfrog.pdf`.

## Problem 14: Make a unified implementation of many schemes

Consider the linear ODE problem  $u'(t) = -a(t)u(t) + b(t)$ ,  $u(0) = I$ . Explicit schemes for this problem can be written in the general form

$$u^{n+1} = \sum_{j=0}^m c_j u^{n-j}, \quad (118)$$

for some choice of  $c_0, \dots, c_m$ . Find expressions for the  $c_j$  coefficients in case of the  $\theta$ -rule, the three-level backward scheme, the Leapfrog scheme, the 2nd-order Runge-Kutta method, and the 3rd-order Adams-Bashforth scheme.

Make a class `ExpDecay` that implements the general updating formula (118). The formula cannot be applied for  $n < m$ , and for those  $n$  values, other schemes must be used. Assume for simplicity that we just repeat Crank-Nicolson steps

until (118) can be used. Use a subclass to specify the list  $c_0, \dots, c_m$  for a particular method, and implement subclasses for all the mentioned schemes. Verify the implementation by testing with a linear solution, which should be exactly reproduced by all methods. Filename: `decay_schemes_oo.py`.

## 9 Applications of exponential decay models

This section presents many mathematical models that all end up with ODEs of the type  $u' = -au + b$ . The applications are taken from biology, finance, and physics, and cover population growth or decay, compound interest and inflation, radioactive decay, cooling of objects, compaction of geological media, pressure variations in the atmosphere, and air resistance on falling or rising bodies.

### 9.1 Scaling

Real applications of a model  $u' = -au + b$  will often involve a lot of parameters in the expressions for  $a$  and  $b$ . It can be quite a challenge to find relevant values of all parameters. In simple problems, however, it turns out that it is not always necessary to estimate all parameters because we can lump them into one or a few *dimensionless* numbers by using a very attractive technique called scaling. It simply means to stretch the  $u$  and  $t$  axis in the present problem - and suddenly all parameters in the problem are lumped one parameter if  $b \neq 0$  and no parameter when  $b = 0$ !

Scaling means that we introduce a new function  $\bar{u}(\bar{t})$ , with

$$\bar{u} = \frac{u - u_m}{u_c}, \quad \bar{t} = \frac{t}{t_c},$$

where  $u_m$  is a characteristic value of  $u$ ,  $u_c$  is a characteristic size of the range of  $u$  values, and  $t_c$  is a characteristic size of the range of  $t$  where  $u$  varies significantly. Choosing  $u_m$ ,  $u_c$ , and  $t_c$  is not always easy and often an art in complicated problems. We just state one choice first:

$$u_c = I, \quad u_m = b/a, \quad t_c = 1/a.$$

Inserting  $u = u_m + u_c \bar{u}$  and  $t = t_c \bar{t}$  in the problem  $u' = -au + b$ , assuming  $a$  and  $b$  are constants, results after some algebra in the *scaled problem*

$$\frac{d\bar{u}}{d\bar{t}} = -\bar{u}, \quad \bar{u}(0) = 1 - \beta,$$

where  $\beta$  is a dimensionless number

$$\beta = \frac{b}{Ia}.$$

That is, only the special combination of  $b/(Ia)$  matters, not what the individual values of  $b$ ,  $a$ , and  $I$  are. Moreover, if  $b = 0$ , the scaled problem is independent of  $a$  and  $I$ ! In practice this means that we can perform one numerical simulation

of the scaled problem and recover the solution of any problem for a given  $a$  and  $I$  by stretching the axis in the plot:  $u = I\bar{u}$  and  $t = \bar{t}/a$ . For  $b \neq 0$ , we simulate the scaled problem for a few  $\beta$  values and recover the physical solution  $u$  by translating and stretching the  $u$  axis and stretching the  $t$  axis.

The scaling breaks down if  $I = 0$ . In that case we may choose  $u_m = 0$ ,  $u_c = b/a$ , and  $t_c = 1/b$ , resulting in a slightly different scaled problem:

$$\frac{d\bar{u}}{d\bar{t}} = 1 - \bar{u}, \quad \bar{u}(0) = 0.$$

As with  $b = 0$ , the case  $I = 0$  has a scaled problem with no physical parameters!

It is common to drop the bars after scaling and write the scaled problem as  $u' = -u$ ,  $u(0) = 1 - \beta$ , or  $u' = 1 - u$ ,  $u(0) = 0$ . Any implementation of the problem  $u' = -au + b$ ,  $u(0) = I$ , can be reused for the scaled problem by setting  $a = 1$ ,  $b = 0$ , and  $I = 1 - \beta$  in the code, if  $I \neq 0$ , or one sets  $a = 1$ ,  $b = 1$ , and  $I = 0$  when the physical  $I$  is zero. Falling bodies in fluids, as described in Section 9.8, involves  $u' = -au + b$  with seven physical parameters. All these vanish in the scaled version of the problem if we start the motion from rest!

## 9.2 Evolution of a population

Let  $N$  be the number of individuals in a population occupying some spatial domain. Despite  $N$  being an integer in this problem, we shall compute with  $N$  as a real number and view  $N(t)$  as a continuous function of time. The basic model assumption is that in a time interval  $\Delta t$  the number of newcomers to the populations (newborns) is proportional to  $N$ , with proportionality constant  $\bar{b}$ . The amount of newcomers will increase the population and result in to

$$N(t + \Delta t) = N(t) + \bar{b}N(t).$$

It is obvious that a long time interval  $\Delta t$  will result in more newcomers and hence a larger  $\bar{b}$ . Therefore, we introduce  $b = \bar{b}/\Delta t$ : the number of newcomers per unit time and per individual. We must then multiply  $b$  by the length of the time interval considered and by the population size to get the total number of new individuals,  $b\Delta tN$ .

If the number of removals from the population (deaths) is also proportional to  $N$ , with proportionality constant  $d\Delta t$ , the population evolves according to

$$N(t + \Delta t) = N(t) + b\Delta tN(t) - d\Delta tN(t).$$

Dividing by  $\Delta t$  and letting  $\Delta t \rightarrow 0$ , we get the ODE

$$N' = (b - d)N, \quad N(0) = N_0. \quad (119)$$

In a population where the death rate ( $d$ ) is larger than the newborn rate ( $b$ ),  $a > 0$ , and the population experiences exponential decay rather than exponential growth.

In some populations there is an immigration of individuals into the spatial domain. With  $I$  individuals coming in per time unit, the equation for the population change becomes

$$N(t + \Delta t) = N(t) + b\Delta t N(t) - d\Delta t N(t) + \Delta t I.$$

The corresponding ODE reads

$$N' = (b - d)N + I, \quad N(0) = N_0. \quad (120)$$

Some simplification arises if we introduce a fractional measure of the population:  $u = N/N_0$  and set  $r = b - d$ . The ODE problem now becomes

$$u' = ru + f, \quad u(0) = 1, \quad (121)$$

where  $f = I/N_0$  measures the net immigration per time unit as the fraction of the initial population. Very often,  $r$  is approximately constant, but  $f$  is usually a function of time.

The growth rate  $r$  of a population decreases if the environment has limited resources. Suppose the environment can sustain at most  $N_{\max}$  individuals. We may then assume that the growth rate approaches zero as  $N$  approaches  $N_{\max}$ , i.e., as  $u$  approaches  $M = N_{\max}/N_0$ . The simplest possible evolution of  $r$  is then a linear function:  $r(t) = r_0(1 - u(t)/M)$ , where  $r_0$  is the initial growth rate when the population is small relative to the maximum size and there is enough resources. Using this  $r(t)$  in (121) results in the *logistic model* for the evolution of a population (assuming for the moment that  $f = 0$ ):

$$u' = r_0(1 - u/M)u, \quad u(0) = 1. \quad (122)$$

Initially,  $u$  will grow at rate  $r_0$ , but the growth will decay as  $u$  approaches  $M$ , and then there is no more change in  $u$ , causing  $u \rightarrow M$  as  $t \rightarrow \infty$ . Note that the logistic equation  $u' = r_0(1 - u/M)u$  is *nonlinear* because of the quadratic term  $-u^2 r_0/M$ .

### 9.3 Compound interest and inflation

Say the annual interest rate is  $r$  percent and that the bank adds the interest once a year to your investment. If  $u^n$  is the investment in year  $n$ , the investment in year  $u^{n+1}$  grows to

$$u^{n+1} = u^n + \frac{r}{100} u^n.$$

In reality, the interest rate is added every day. We therefore introduce a parameter  $m$  for the number of periods per year when the interest is added. If  $n$  counts the periods, we have the fundamental model for compound interest:

$$u^{n+1} = u^n + \frac{r}{100m} u^n. \quad (123)$$

This model is a *difference equation*, but it can be transformed to a continuous differential equation through a limit process. The first step is to derive a formula for the growth of the investment over a time  $t$ . Starting with an investment  $u^0$ , and assuming that  $r$  is constant in time, we get

$$\begin{aligned} u^{n+1} &= \left(1 + \frac{r}{100m}\right) u^n \\ &= \left(1 + \frac{r}{100m}\right)^2 u^{n-1} \\ &\vdots \\ &= \left(1 + \frac{r}{100m}\right)^{n+1} u^0 \end{aligned}$$

Introducing time  $t$ , which here is a real-numbered counter for years, we have that  $n = mt$ , so we can write

$$u^{mt} = \left(1 + \frac{r}{100m}\right)^{mt} u^0.$$

The second step is to assume *continuous compounding*, meaning that the interest is added continuously. This implies  $m \rightarrow \infty$ , and in the limit one gets the formula

$$u(t) = u_0 e^{rt/100}, \quad (124)$$

which is nothing but the solution of the ODE problem

$$u' = \frac{r}{100} u, \quad u(0) = u_0. \quad (125)$$

This is then taken as the ODE model for compound interest if  $r > 0$ . However, the reasoning applies equally well to inflation, which is just the case  $r < 0$ . One may also take the  $r$  in (125) as the net growth of an investemt, where  $r$  takes both compound interest and inflation into account. Note that for real applications we must use a time-dependent  $r$  in (125).

Introducing  $a = \frac{r}{100}$ , continuous inflation of an initial fortune  $I$  is then a process exhibiting exponential decay according to

$$u' = -au, \quad u(0) = I.$$

## 9.4 Radioactive decay

An atomic nucleus of an unstable atom may lose energy by emitting ionizing particles and thereby be transformed to a nucleus with a different number of protons and neutrons. This process is known as [radioactive decay](#). Actually, the process is stochastic when viewed for a single atom, because it is impossible to predict exactly when a particular atom emits a particle. Nevertheless, with a large number of atoms,  $N$ , one may view the process as deterministic and compute the mean behavior of the decay. Below we reason intuitively about an ODE for the mean behavior. Thereafter, we show mathematically that a detailed stochastic model for single atoms leads the same mean behavior.



**Deterministic model.** Suppose at time  $t$ , the number of the original atom type is  $N(t)$ . A basic model assumption is that the transformation of the atoms of the original type in a small time interval  $\Delta t$  is proportional to  $N$ , so that

$$N(t + \Delta t) = N(t) - a\Delta t N(t),$$

where  $a > 0$  is a constant. Introducing  $u = N(t)/N(0)$ , dividing by  $\Delta t$  and letting  $\Delta t \rightarrow 0$  gives the following ODE:

$$u' = -au, \quad u(0) = 1. \quad (126)$$

The parameter  $a$  can for a given nucleus be expressed through the *half-life*  $t_{1/2}$ , which is the time taken for the decay to reduce the initial amount by one half, i.e.,  $u(t_{1/2}) = 0.5$ . With  $u(t) = e^{-at}$ , we get  $t_{1/2} = a^{-1} \ln 2$  or  $a = \ln 2 / t_{1/2}$ .

**Stochastic model.** We have originally  $N_0$  atoms. Each atom may have decayed or survived at a particular time  $t$ . We want to count how many original atoms that are left, i.e., how many atoms that have survived. The survival of a single atom at time  $t$  is a random event. Since there are only two outcomes, survival or decay, we have a [Bernoulli trial](#). Let  $p$  be the probability of survival (implying that the probability of decay is  $1 - p$ ). If each atom survives independently of the others, and the probability of survival is the same for every atom, we have  $N_0$  statistically Bernoulli trials, known as a *binomial experiment* from probability theory. The probability  $P(N)$  that  $N$  out of the  $N_0$  atoms have survived at time  $t$  is then given by the famous *binomial distribution*

$$P(N) = \frac{N_0!}{N!(N_0 - N)!} p^N (1 - p)^{N_0 - N}.$$

The mean (or expected) value  $E[P]$  of  $P(N)$  is known to be  $N_0 p$ .

It remains to estimate  $p$ . Let the interval  $[0, t]$  be divided into  $m$  small subintervals of length  $\Delta t$ . We make the assumption that the probability of decay of a single atom in an interval of length  $\Delta t$  is  $\tilde{p}$ , and that this probability is proportional to  $\Delta t$ :  $\tilde{p} = \lambda \Delta t$  (it sounds natural that the probability of decay increases with  $\Delta t$ ). The corresponding probability of survival is  $1 - \lambda \Delta t$ . Believing that  $\lambda$  is independent of time, we have, for each interval of length  $\Delta t$ , a Bernoulli trial: the atom either survives or decays in that interval. Now,  $p$  should be the probability that the atom survives in all the intervals, i.e., that we have  $m$  successful Bernoulli trials in a row and therefore

$$p = (1 - \lambda \Delta t)^m.$$

The expected number of atoms of the original type at time  $t$  is

$$E[P] = N_0 p = N_0 (1 - \lambda \Delta t)^m, \quad m = t / \Delta t. \quad (127)$$

To see the relation between the two types of Bernoulli trials and the ODE above, we go to the limit  $\Delta t \rightarrow 0$ ,  $m \rightarrow \infty$ . One can show that

$$p = \lim_{m \rightarrow \infty} (1 - \lambda \Delta t)^m = \lim_{m \rightarrow \infty} \left(1 - \lambda \frac{t}{m}\right)^m = e^{-\lambda t}$$

This is the famous exponential waiting time (or arrival time) distribution for a Poisson process in probability theory (obtained here, as often done, as the limit of a binomial experiment). The probability of decay,  $1 - e^{-\lambda t}$ , follows an [exponential distribution](#). The limit means that  $m$  is very large, hence  $\Delta t$  is very small, and  $\tilde{p} = \lambda \Delta t$  is very small since the intensity of the events,  $\lambda$ , is assumed finite. This situation corresponds to a very small probability that an atom will decay in a very short time interval, which is a reasonable model. The same model occurs in lots of different applications, e.g., when waiting for a taxi, or when finding defects along a rope.

**Relation between stochastic and deterministic models.** With  $p = e^{-\lambda t}$  we get the expected number of original atoms at  $t$  as  $N_0 p = N_0 e^{-\lambda t}$ , which is exactly the solution of the ODE model  $N' = -\lambda N$ . This gives also an interpretation of  $a$  via  $\lambda$  or vice versa. Our important finding here is that the ODE model captures the mean behavior of the underlying stochastic model. This is, however, not always the common relation between microscopic stochastic models and macroscopic "averaged" models.

Also of interest is to see that a Forward Euler discretization of  $N' = -\lambda N$ ,  $N(0) = N_0$ , gives  $N^m = N_0(1 - \lambda \Delta t)^m$  at time  $t_m = m \Delta t$ , which is exactly the expected value of the stochastic experiment with  $N_0$  atoms and  $m$  small intervals of length  $\Delta t$ , where each atom can decay with probability  $\lambda \Delta t$  in an interval.

A fundamental question is how accurate the ODE model is. The underlying stochastic model fluctuates around its expected value. A measure of the fluctuations is the standard deviation of the binomial experiment with  $N_0$  atoms, which can be shown to be  $\text{Std}[P] = \sqrt{N_0 p(1 - p)}$ . Compared to the size of the expectation, we get the normalized standard deviation

$$\frac{\sqrt{\text{Var}[P]}}{\text{E}[P]} = N_0^{-1/2} \sqrt{p^{-1} - 1} = N_0^{-1/2} \sqrt{(1 - e^{-\lambda t})^{-1} - 1} \approx (N_0 \lambda t)^{-1/2},$$

showing that the normalized fluctuations are very small if  $N_0$  is very large, which is usually the case.

## 9.5 Newton's law of cooling

When a body at some temperature is placed in a cooling environment, experience shows that the temperature falls rapidly in the beginning, and then the changes in temperature levels off until the body's temperature equals that of the surroundings. Newton carried out some experiments on cooling hot iron and found that the temperature evolved as a "geometric progression at times in arithmetic progression", meaning that the temperature decayed exponentially.

Later, this result was formulated as a differential equation: the rate of change of the temperature in a body is proportional to the temperature difference between the body and its surroundings. This statement is known as *Newton's law of cooling*, which can be mathematically expressed as

$$\frac{dT}{dt} = -k(T - T_s), \quad (128)$$

where  $T$  is the temperature of the body,  $T_s$  is the temperature of the surroundings, which may be time-dependent,  $t$  is time, and  $k$  is a positive constant. Equation (146) is primarily viewed as an empirical law, valid when heat is efficiently convected away from the surface of the body by a flowing fluid such as air at constant temperature  $T_s$ . The *heat transfer coefficient*  $k$  reflects the transfer of heat from the body to the surroundings and must be determined from physical experiments.

The cooling law (146) needs an initial condition  $T(0) = T_0$ .

## 9.6 Decay of atmospheric pressure with altitude

Vertical equilibrium of air in the atmosphere is governed by the equation

$$\frac{dp}{dz} = -\varrho g. \quad (129)$$

Here,  $p(z)$  is the air pressure,  $\varrho$  is the density of air, and  $g = 9.807 \text{ m/s}^2$  is a standard value of the acceleration of gravity. (Equation (129) follows directly from the general Navier-Stokes equations for fluid motion, with the assumption that the air does not move.)

The pressure is related to density and temperature through the ideal gas law

$$\varrho = \frac{Mp}{R^*T}, \quad (130)$$

where  $M$  is the molar mass of the Earth's air (0.029 kg/mol),  $R^*$  is the universal gas constant (8.314 Nm/(mol K)), and  $T$  is the temperature. All variables  $p$ ,  $\varrho$ , and  $T$  vary with the height  $z$ . Inserting (130) in (129) results in an ODE with a variable coefficient:

$$\frac{dp}{dz} = -\frac{Mg}{R^*T(z)}p. \quad (131)$$

**Multiple atmospheric layers.** The atmosphere can be approximately modeled by seven layers. In each layer, (131) is applied with a linear temperature of the form

$$T(z) = \bar{T}_i + L_i(z - h_i),$$

where  $z = h_i$  denotes the bottom of layer number  $i$ , having temperature  $\bar{T}_i$ , and  $L_i$  is a constant in layer number  $i$ . The table below lists  $h_i$  (m),  $\bar{T}_i$  (K), and  $L_i$  (K/m) for the layers  $i = 0, \dots, 6$ .

$i$	$h_i$	$\bar{T}_i$	$L_i$
0	0	288	-0.0065
1	11,000	216	0.0
2	20,000	216	0.001
3	32,000	228	0.0028
4	47,000	270	0.0
5	51,000	270	-0.0028
6	71,000	214	-0.002

For implementation it might be convenient to write (131) on the form

$$\frac{dp}{dz} = -\frac{Mg}{R^*(\bar{T}(z) + L(z)(z - h(z)))}p, \quad (132)$$

where  $\bar{T}(z)$ ,  $L(z)$ , and  $h(z)$  are piecewise constant functions with values given in the table. The value of the pressure at the sea level  $z = 0$ ,  $p_0 = p(0)$ , is 101325 Pa.

**Simplification:**  $L = 0$ . One commonly used simplification is to assume that the temperature is constant within each layer. This means that  $L = 0$ .

**Simplification: one-layer model.** Another commonly used approximation is to work with one layer instead of seven. This **one-layer model** is based on  $T(z) = T_0 - Lz$ , with sea level standard temperature  $T_0 = 288$  K and temperature lapse rate  $L = 0.0065$  K/m.

## 9.7 Compaction of sediments

Sediments, originally made from materials like sand and mud, get compacted through geological time by the weight of new material that is deposited on the sea bottom. The porosity  $\phi$  of the sediments tells how much void (fluid) space there is between the sand and mud grains. The porosity reduces with depth because the weight of the sediments above and causes the void space to shrink and thereby increase the compaction.

A typical assumption is that the change in  $\phi$  at some depth  $z$  is negatively proportional to  $\phi$ . This assumption leads to the differential equation problem

$$\frac{d\phi}{dz} = -c\phi, \quad \phi(0) = \phi_0, \quad (133)$$

where the  $z$  axis points downwards,  $z = 0$  is the surface with known porosity, and  $c > 0$  is a constant.

The upper part of the Earth's crust consists of many geological layers stacked on top of each other, as indicated in Figure 18. The model (133) can be applied for each layer. In layer number  $i$ , we have the unknown porosity function  $\phi_i(z)$  fulfilling  $\phi_i'(z) = -c_i z$ , since the constant  $c$  in the model (133) depends on the type of sediment in the layer. Alternatively, we can use (133) to describe the

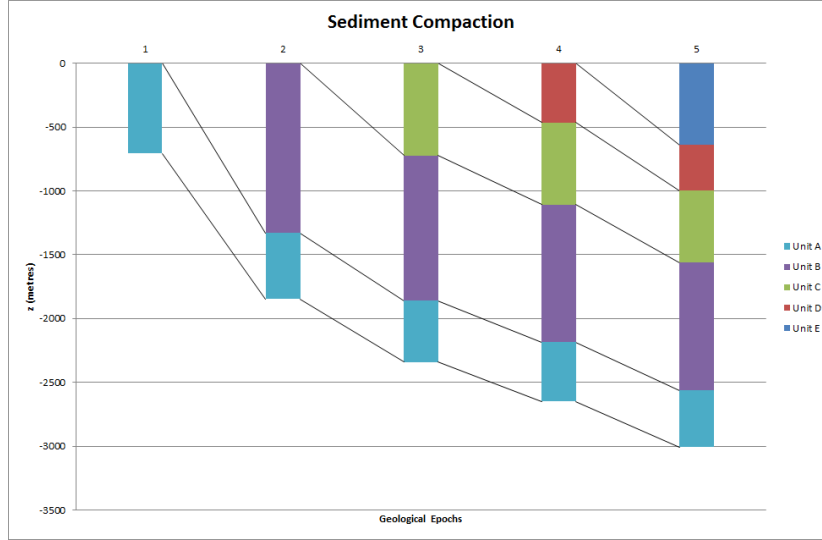


Figure 18: Illustration of the compaction of geological layers (with different colors) through time.

porosity through all layers if  $c$  is taken as a piecewise constant function of  $z$ , equal to  $c_i$  in layer  $i$ . From the figure we see that new layers of sediments are deposited on top of older ones as time progresses. The compaction, as measured by  $\phi$ , is rapid in the beginning and then decreases (exponentially) with depth in each layer.

When we drill a well at present time through the right-most column of sediments in Figure 18, we can measure the thickness of the sediment in (say) the bottom layer. Let  $L_1$  be this thickness. Assuming that the volume of sediment remains constant through time, we have that the initial volume,  $\int_0^{L_{1,0}} \phi_1 dz$ , must equal the volume seen today,  $\int_{\ell-L_1}^{\ell} \phi_1 dz$ , where  $\ell$  is the depth of the bottom of the sediment in the present day configuration. After having solved for  $\phi_1$  as a function of  $z$ , we can then find the original thickness  $L_{1,0}$  of the sediment from the equation

$$\int_0^{L_{1,0}} \phi_1 dz = \int_{\ell-L_1}^{\ell} \phi_1 dz.$$

In hydrocarbon exploration it is important to know  $L_{1,0}$  and the compaction history of the various layers of sediments.

## 9.8 Vertical motion of a body in a viscous fluid

A body moving vertically through a fluid (liquid or gas) is subject to three different types of forces: the gravity force, the drag force, and the buoyancy force.

**Overview of forces.** The gravity force is  $F_g = -mg$ , where  $m$  is the mass of the body and  $g$  is the acceleration of gravity. The uplift or buoyancy force ("Archimedes force") is  $F_b = \varrho gV$ , where  $\varrho$  is the density of the fluid and  $V$  is the volume of the body. Forces and other quantities are taken as positive in the upward direction.

The drag force is of two types, depending on the Reynolds number

$$\text{Re} = \frac{\varrho d|v|}{\mu}, \quad (134)$$

where  $d$  is the diameter of the body in the direction perpendicular to the flow,  $v$  is the velocity of the body, and  $\mu$  is the dynamic viscosity of the fluid. When  $\text{Re} < 1$ , the drag force is fairly well modeled by the so-called Stokes' drag, which for a spherical body of diameter  $d$  reads

$$F_d^{(S)} = -3\pi d\mu v. \quad (135)$$

For large  $\text{Re}$ , typically  $\text{Re} > 10^3$ , the drag force is quadratic in the velocity:

$$F_d^{(q)} = -\frac{1}{2}C_D\varrho A|v|v, \quad (136)$$

where  $C_D$  is a dimensionless drag coefficient depending on the body's shape, and  $A$  is the cross-sectional area as produced by a cut plane, perpendicular to the motion, through the thickest part of the body. The superscripts  $^{(q)}$  and  $^{(S)}$  in  $F_d^{(S)}$  and  $F_d^{(q)}$  indicate Stokes drag and quadratic drag, respectively.

**Equation of motion.** All the mentioned forces act in the vertical direction. Newton's second law of motion applied to the body says that the sum of these forces must equal the mass of the body times its acceleration  $a$  in the vertical direction.

$$ma = F_g + F_d^{(S)} + F_b.$$

Here we have chosen to model the fluid resistance by the Stokes drag. Inserting the expressions for the forces yields

$$ma = -mg - 3\pi d\mu v + \varrho gV.$$

The unknowns here are  $v$  and  $a$ , i.e., we have two unknowns but only one equation. From kinematics in physics we know that the acceleration is the time derivative of the velocity:  $a = dv/dt$ . This is our second equation. We can easily eliminate  $a$  and get a single differential equation for  $v$ :

$$m \frac{dv}{dt} = -mg - 3\pi d\mu v + \varrho gV.$$

A small rewrite of this equation is handy: We express  $m$  as  $\varrho_b V$ , where  $\varrho_b$  is the density of the body, and we divide by the mass to get

$$v'(t) = -\frac{3\pi d\mu}{\varrho_b V}v + g\left(\frac{\varrho}{\varrho_b} - 1\right). \quad (137)$$

We may introduce the constants

$$a = \frac{3\pi d\mu}{\varrho_b V}, \quad b = g\left(\frac{\varrho}{\varrho_b} - 1\right), \quad (138)$$

so that the structure of the differential equation becomes obvious:

$$v'(t) = -av(t) + b. \quad (139)$$

The corresponding initial condition is  $v(0) = v_0$  for some prescribed starting velocity  $v_0$ .

This derivation can be repeated with the quadratic drag force  $F_d^{(q)}$ , leading to the result

$$v'(t) = -\frac{1}{2}C_D \frac{\varrho A}{\varrho_b V}|v|v + g\left(\frac{\varrho}{\varrho_b} - 1\right). \quad (140)$$

Defining

$$a = \frac{1}{2}C_D \frac{\varrho A}{\varrho_b V}, \quad (141)$$

and  $b$  as above, we can write (140) as

$$v'(t) = -a|v|v + b. \quad (142)$$

**Terminal velocity.** An interesting aspect of (139) and (142) is whether  $v$  will approach a final constant value, the so-called *terminal velocity*  $v_T$ , as  $t \rightarrow \infty$ . A constant  $v$  means that  $v'(t) \rightarrow 0$  as  $t \rightarrow \infty$  and therefore the terminal velocity  $v_T$  solves

$$0 = -av_T + b$$

and

$$0 = -a|v_T|v_T + b.$$

The former equation implies  $v_T = b/a$ , while the latter has solutions  $v_T = -\sqrt{|b|/a}$  for a falling body ( $v_T < 0$ ) and  $v_T = \sqrt{|b|/a}$  for a rising body ( $v_T > 0$ ).

**A Crank-Nicolson scheme.** Both governing equations, the Stokes' drag model (139) and the quadratic drag model (142), can be readily solved by the Forward Euler scheme. For higher accuracy one can use the Crank-Nicolson method, but a straightforward application this method results a nonlinear equation in the new unknown value  $v^{n+1}$  when applied to (142):

$$\frac{v^{n+1} - v^n}{\Delta t} = -a \frac{1}{2} (|v^{n+1}|v^{n+1} + |v^n|v^n) + b. \quad (143)$$

However, instead of approximating the term  $-|v|v$  by an arithmetic average, we can use a *geometric mean*:

$$(|v|v)^{n+\frac{1}{2}} \approx |v^n|v^{n+1}. \quad (144)$$

The error is of second order in  $\Delta t$ , just as for the arithmetic average and the centered finite difference approximation in (143). With this approximation trick, the discrete equation

$$\frac{v^{n+1} - v^n}{\Delta t} = -a|v^n|v^{n+1} + b$$

becomes a *linear* equation in  $v^{n+1}$ , and we can therefore easily solve for  $v^{n+1}$ :

$$v^{n+1} = \frac{v^n + \Delta t b^{n+\frac{1}{2}}}{1 + \Delta t a^{n+\frac{1}{2}} |v^n|}. \quad (145)$$

**Physical data.** Suitable values of  $\mu$  are  $1.8 \cdot 10^{-5}$  Pa s for air and  $8.9 \cdot 10^{-4}$  Pa s for water. Densities can be taken as  $1.2 \text{ kg/m}^3$  for air and as  $1.0 \cdot 10^3 \text{ kg/m}^3$  for water. For considerable vertical displacement in the atmosphere one should take into account that the density of air varies with the altitude, see Section 9.6. One possible density variation arises from the one-layer model in the mentioned section.

Any density variation makes  $b$  time dependent and we need  $b^{n+\frac{1}{2}}$  in (145). To compute the density that enters  $b^{n+\frac{1}{2}}$  we must also compute the vertical position  $z(t)$  of the body. Since  $v = dz/dt$ , we can use a centered difference approximation:

$$\frac{z^{n+\frac{1}{2}} - z^{n-\frac{1}{2}}}{\Delta t} = v^n \quad \Rightarrow \quad z^{n+\frac{1}{2}} = z^{n-\frac{1}{2}} + \Delta t v^n.$$

This  $z^{n+\frac{1}{2}}$  is used in the expression for  $b$  to compute  $\varrho(z^{n+\frac{1}{2}})$  and then  $b^{n+\frac{1}{2}}$ .

The **drag coefficient**  $C_D$  depends heavily on the shape of the body. Some values are: 0.45 for a sphere, 0.42 for a semi-sphere, 1.05 for a cube, 0.82 for a long cylinder (when the center axis is in the vertical direction), 0.75 for a rocket, 1.0-1.3 for a man in upright position, 1.3 for a flat plate perpendicular to the flow, and 0.04 for a streamlined, droplet-like body.

**Verification.** To verify the program, one may assume a heavy body in air such that the  $F_b$  force can be neglected, and further assume a small velocity such that the air resistance  $F_d$  can also be neglected. This can be obtained by setting  $\mu$  and  $\varrho$  to zero. The motion then leads to the velocity  $v(t) = v_0 - gt$ , which is linear in  $t$  and therefore should be reproduced to machine precision (say tolerance  $10^{-15}$ ) by any implementation based on the Crank-Nicolson or Forward Euler schemes.



Another verification, but not as powerful as the one above, can be based on computing the terminal velocity and comparing with the exact expressions. The advantage of this verification is that we can also the test situation  $\varrho \neq 0$ .

As always, the method of manufactured solutions can be applied to test the implementation of all terms in the governing equation, but the solution then has no physical relevance in general.

**Scaling.** Applying scaling, as described in Section 9.1, will for the linear case reduce the need to estimate values for seven parameters down to choosing one value of a single dimensionless parameter

$$\beta = \frac{\varrho_b g V \left( \frac{\varrho}{\varrho_b} - 1 \right)}{3\pi d \mu I},$$

provided  $I \neq 0$ . If the motion starts from rest,  $I = 0$ , the scaled problem  $\bar{u}' = 1 - \bar{u}$ ,  $\bar{u}(0) = 0$ , has no need for estimating physical parameters. This means that there is a single universal solution to the problem of a falling body starting from rest:  $\bar{u}(\bar{t}) = 1 - e^{-\bar{t}}$ . All real physical cases correspond to stretching the  $\bar{t}$  axis and the  $\bar{u}$  axis in this dimensionless solution. More precisely, the physical velocity  $u(t)$  is related to the dimensionless velocity  $\bar{u}(\bar{t})$  through

$$u = \frac{\varrho_b g V \left( \frac{\varrho}{\varrho_b} - 1 \right)}{3\pi d \mu} \bar{u}(\bar{t}/(g(\varrho/\varrho_b - 1))).$$

## 9.9 Decay ODEs from solving a PDE by Fourier expansions

Suppose we have a partial differential equation

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2} + f(x, t),$$

with boundary conditions  $u(0, t) = u(L, t) = 0$  and initial condition  $u(x, 0) = I(x)$ . One may express the solution as

$$u(x, t) = \sum_{k=1}^m A_k(t) e^{ikx\pi/L},$$

for appropriate unknown functions  $A_k$ ,  $k = 1, \dots, m$ . We use the complex exponential  $e^{ikx\pi/L}$  for easy algebra, but the physical  $u$  is taken as the real part of any complex expression. Note that the expansion in terms of  $e^{ikx\pi/L}$  is compatible with the boundary conditions: all functions  $e^{ikx\pi/L}$  vanish for  $x = 0$  and  $x = L$ . Suppose we can express  $I(x)$  as

$$I(x) = \sum_{k=1}^m I_k e^{ikx\pi/L}.$$

Such an expansion can be computed by well-known Fourier expansion techniques, but the details are not important here. Also, suppose we can express the given  $f(x, t)$  as

$$f(x, t) = \sum_{k=1}^m b_k(t) e^{ikx\pi/L}.$$

Inserting the expansions for  $u$  and  $f$  in the differential equations demands that all terms corresponding to a given  $k$  must be equal. The calculations results in the follow system of ODEs:

$$A'_k(t) = -\alpha \frac{k^2 \pi^2}{L^2} + b_k(t), \quad k = 1, \dots, m.$$

From the initial condition

$$u(x, 0) = \sum_k A_k(0) e^{ikx\pi/L} = I(x) = \sum_k I_k e^{(ikx\pi/L)},$$

it follows that  $A_k(0) = I_k$ ,  $k = 1, \dots, m$ . We then have  $m$  equations of the form  $A'_k = -a_k A_k + b_k$ ,  $A_k(0) = I_k$ , for appropriate definitions of  $a$  and  $b$ . These ODE problems independent each other such that we can solve one problem at a time. The outline technique is a quite common approach for solving partial differential equations.

**Remark.** Since  $a_k$  depends on  $k$  and the stability of the Forward Euler scheme demands  $a_k \Delta t \leq 1$ , we get that  $\Delta t \leq \alpha^{-1} L^2 \pi^{-2} k^{-2}$ . Usually, quite large  $k$  values are needed to accurately represent the given functions  $I$  and  $f$  and then  $\Delta t$  needs to be very small for these large values of  $k$ . Therefore, the Crank-Nicolson and Backward Euler schemes, which allow larger  $\Delta t$  without any growth in the solutions, are more popular choices when creating time-stepping algorithms for partial differential equations of the type considered in this example.

## 10 Exercises

### Exercise 15: Derive schemes for Newton's law of cooling

Show in detail how we can apply the ideas of the Forward Euler, Backward Euler, and Crank-Nicolson discretizations to derive explicit computational formulas for new temperature values in Newton's law of cooling (see Section 9.5):

$$\frac{dT}{dt} = -k(T - T_s(t)), \quad T(0) = T_0. \quad (146)$$

Here,  $T$  is the temperature of the body,  $T_s(t)$  is the temperature of the surroundings,  $t$  is time,  $k$  is the heat transfer coefficient, and  $T_0$  is the initial temperature of the body. Summarize the discretizations in a  $\theta$ -rule such that you can get the three schemes from a single formula by varying the  $\theta$  parameter. Filename: `schemes_cooling.pdf`.

## Exercise 16: Implement schemes for Newton's law of cooling

The goal of this exercise is to implement the schemes from Exercise 15 and investigate several approaches for verifying the implementation.

- a) Implement the  $\theta$ -rule from Exercise 15 in a function

```
cooling(T0, k, T_s, t_end, dt, theta=0.5)
```

where  $T_0$  is the initial temperature,  $k$  is the heat transfer coefficient,  $T_s$  is a function of  $t$  for the temperature of the surroundings,  $t_{\text{end}}$  is the end time of the simulation,  $dt$  is the time step, and  $\theta$  corresponds to  $\theta$ . The `cooling` function should return the temperature as an array  $T$  of values at the mesh points and the time mesh  $t$ .

- b) In the case  $\lim_{t \rightarrow \infty} T_s(t) = C = \text{const}$ , explain why  $T(t) \rightarrow C$ . Construct an example where you can illustrate this property in a plot. Implement a corresponding test function that checks the correctness of the asymptotic value of the solution.

- c) A piecewise constant surrounding temperature,

$$T_s(t) = \begin{cases} C_0, & 0 \leq t \leq t^* \\ C_1, & t > t^*, \end{cases}$$

corresponds to a sudden change in the environment at  $t = t^*$ . Choose  $C_1 = 2T_0$ ,  $C_2 = \frac{1}{2}T_0$ , and  $t^* = 4/k$ . Plot the solution and explain why it seems physically reasonable.

- d) Find a value of  $\Delta t$  such that you can illustrate stability problems of the Crank-Nicolson scheme in the test case from c).

- e) Find an expression for the exact solution of (146). Construct a test case and compare the numerical and exact solution in a plot. Find a value of the time step  $\Delta t$  such that the two solution curves cannot (visually) be distinguished from each other. Many scientists will claim that such a plot provides evidence for a correct implementation, but point out why there still may be errors in the code. Can you introduce bugs in the `cooling` function and still achieve visually coinciding curves?

**Hint.** The exact solution can be derived by multiplying (146) by the integrating factor  $e^{kt}$ .

- f) Implement a test function for checking that the solution returned by the `cooling` function is identical to the exact numerical solution of the problem (to machine precision) when  $T_s$  is constant.

**Hint.** The exact solution of the discrete equations in the case  $T_s$  is a constant can be found by introducing  $u = T - T_s$  to get a problem  $u' = -ku$ ,  $u(0) = T_0 - T_s$ . The solution of the discrete equations is then of the form  $u^n = (T_0 - T_s)A^n$  for some amplification factor  $A$ . This approach can be used to derive an expression for  $T^n$ .

Filename: `cooling.py`.

### Exercise 17: Find time of murder from body temperature

A detective measures the temperature of a dead body to be 26.7 C at 2 pm. One hour later the temperature is 25.8 C. The question is when death occurred.

Assume that Newton's law of cooling (146) is an appropriate mathematical model for the evolution of the temperature in the body. First, determine  $k$  in (146) by formulating a Forward Euler approximation with one time step from time 2 am to time 3 am, where knowing the two temperatures allows for finding  $k$ . Assume the temperature in the air to be 20 C. Thereafter, simulate the temperature evolution from the time of murder, taken as  $t = 0$ , when  $T = 37$  C, until the temperature reaches 25.8 C. The corresponding time allows for answering when death occurred. Filename: `detective.py`.

### Exercise 18: Simulate an oscillating cooling process

The surrounding temperature  $T_s$  in Newton's law of cooling (146) may vary in time. Assume that the variations are periodic with period  $P$  and amplitude  $a$  around a constant mean temperature  $T_m$ :

$$T_s(t) = T_m + a \sin\left(\frac{2\pi}{P}t\right). \quad (147)$$

Simulate a process with the following data:  $k = 20 \text{ min}^{-1}$ ,  $T(0) = 5$  C,  $T_m = 25$  C,  $a = 2.5$  C, and  $P = 1$  h. Also experiment with  $P = 10$  min and  $P = 3$  h. Plot  $T$  and  $T_s$  in the same plot. Filename: `osc_cooling.py`.

### Exercise 19: Radioactive decay of Carbon-14

The [Carbon-14](#) isotope, whose radioactive decay is used extensively in dating organic material that is tens of thousands of years old, has a half-life of 5,730 years. Determine the age of an organic material that contains 8.4 percent of its initial amount of Carbon-14. Use a time unit of 1 year in the computations. The uncertainty in the half time of Carbon-14 is  $\pm 40$  years. What is the corresponding uncertainty in the estimate of the age?

**Hint.** Use simulations with  $5,730 \pm 40$  y as input and find the corresponding interval for the result.

Filename: `carbon14.py`.

## Exercise 20: Simulate stochastic radioactive decay

The purpose of this exercise is to implement the stochastic model described in Section 9.4 and show that its mean behavior approximates the solution of the corresponding ODE model.

The simulation goes on for a time interval  $[0, T]$  divided into  $N_t$  intervals of length  $\Delta t$ . We start with  $N_0$  atoms. In some time interval, we have  $N$  atoms that have survived. Simulate  $N$  Bernoulli trials with probability  $\lambda \Delta t$  in this interval by drawing  $N$  random numbers, each being 0 (survival) or 1 (decay), where the probability of getting 1 is  $\lambda \Delta t$ . We are interested in the number of decays,  $d$ , and the number of survived atoms in the next interval is then  $N - d$ . The Bernoulli trials are simulated by drawing  $N$  uniformly distributed real numbers on  $[0, 1]$  and saying that 1 corresponds to a value less than  $\lambda \Delta t$ :

```
# Given lambda_, dt, N
import numpy as np
uniform = np.random.uniform(N)
Bernoulli_trials = np.asarray(uniform < lambda_*dt, dtype=np.int)
d = Bernoulli_trials.size
```

Observe that `uniform < lambda_*dt` is a boolean array whose true and false values become 1 and 0, respectively, when converted to an integer array.

Repeat the simulation over  $[0, T]$  a large number of times, compute the average value of  $N$  in each interval, and compare with the solution of the corresponding ODE model. Filename: `stochastic_decay.py`.

## Exercise 21: Radioactive decay of two substances

Consider two radioactive substances A and B. The nuclei in substance A decay to form nuclei of type B with a half-life  $A_{1/2}$ , while substance B decay to form type A nuclei with a half-life  $B_{1/2}$ . Letting  $u_A$  and  $u_B$  be the fractions of the initial amount of material in substance A and B, respectively, the following system of ODEs governs the evolution of  $u_A(t)$  and  $u_B(t)$ :

$$\frac{1}{\ln 2} u'_A = u_B/B_{1/2} - u_A/A_{1/2}, \quad (148)$$

$$\frac{1}{\ln 2} u'_B = u_A/A_{1/2} - u_B/B_{1/2}, \quad (149)$$

with  $u_A(0) = u_B(0) = 1$ .

Make a simulation program that solves for  $u_A(t)$  and  $u_B(t)$ . Verify the implementation by computing analytically the limiting values of  $u_A$  and  $u_B$  as  $t \rightarrow \infty$  (assume  $u'_A, u'_B \rightarrow 0$ ) and comparing these with those obtained numerically.

Run the program for the case of  $A_{1/2} = 10$  minutes and  $B_{1/2} = 50$  minutes. Use a time unit of 1 minute. Plot  $u_A$  and  $u_B$  versus time in the same plot. Filename: `radioactive_decay_2subst.py`.

## Exercise 22: Simulate the pressure drop in the atmosphere

We consider the models for atmospheric pressure in Section 9.6. Make a program with three functions,

- one computing the pressure  $p(z)$  using a seven-layer model and varying  $L$ ,
- one computing  $p(z)$  using a seven-layer model, but with constant temperature in each layer, and
- one computing  $p(z)$  based on the one-layer model.

How can these implementations be verified? Should ease of verification impact how you code the functions? Compare the three models in a plot. Filename: `atmospheric_pressure.py`.

## Exercise 23: Make a program for vertical motion in a fluid

Implement the Stokes' drag model (137) and the quadratic drag model (140) from Section 9.8, using the Crank-Nicolson scheme and a geometric mean for  $|v|v$  as explained, and assume constant fluid density. At each time level, compute the Reynolds number  $Re$  and choose the Stokes' drag model if  $Re < 1$  and the quadratic drag model otherwise.

The computation of the numerical solution should take place either in a stand-alone function or in a solver class that looks up a problem class for physical data. Create a module and equip it with `pytest/nose` compatible test functions for automatically verifying the code.

Verification tests can be based on

- the terminal velocity (see Section 9.8),
- the exact solution when the drag force is neglected (see Section 9.8),
- the method of manufactured solutions (see Section 6.5) combined with computing convergence rates (see Section 6.6).

Use, e.g., a quadratic polynomial for the velocity in the method of manufactured solutions. The expected error is  $\mathcal{O}(\Delta t^2)$  from the centered finite difference approximation and the geometric mean approximation for  $|v|v$ .

A solution that is linear in  $t$  will also be an exact solution of the discrete equations in many problems. Show that this is true for linear drag (by adding a source term that depends on  $t$ ), but not for quadratic drag because of the geometric mean approximation. Use the method of manufactured solutions to add a source term *in the discrete equations for quadratic drag* such that a linear function of  $t$  is a solution. Add a test function for checking that the linear function is reproduced to machine precision in the case of both linear and quadratic drag.

Apply the software to a case where a ball rises in water. The buoyancy force is here the driving force, but the drag will be significant and balance the other

forces after a short time. A soccer ball has radius 11 cm and mass 0.43 kg. Start the motion from rest, set the density of water,  $\rho$ , to  $1000 \text{ kg/m}^3$ , set the dynamic viscosity,  $\mu$ , to  $10^{-3} \text{ Pa s}$ , and use a drag coefficient for a sphere: 0.45. Plot the velocity of the rising ball. Filename: `vertical_motion.py`.

## Project 24: Simulate parachuting

The aim of this project is to develop a general solver for the vertical motion of a body with quadratic air drag, verify the solver, apply the solver to a skydiver in free fall, and finally apply the solver to a complete parachute jump.

All the pieces of software implemented in this project should be realized as Python functions and/or classes and collected in one module.

**a)** Set up the differential equation problem that governs the velocity of the motion. The parachute jumper is subject to the gravity force and a quadratic drag force. Assume constant density. Add an extra source term be used for program verification. Identify the input data to the problem.

**b)** Make a Python module for computing the velocity of the motion. Also equip the module with functionality for plotting the velocity.

**Hint 1.** Use the Crank-Nicolson scheme with a geometric mean of  $|v|v$  in time to linearize the equation of motion with quadratic drag.

**Hint 2.** You can either use functions or classes for implementation. If you choose functions, make a function `solver` that takes all the input data in the problem as arguments and that returns the velocity (as a mesh function) and the time mesh. In case of a class-based implementation, introduce a problem class with the physical data and a solver class with the numerical data and a `solve` method that stores the velocity and the mesh in the class.

Allow for a time-dependent area and drag coefficient in the formula for the drag force.

**c)** Show that a linear function of  $t$  does not fulfill the discrete equations because of the geometric mean approximation used for the quadratic drag term. Fit a source term, as in the method of manufactured solutions, such that a linear function of  $t$  is a solution of the discrete equations. Make a test function to check that this solution is reproduced to machine precision.

**d)** The expected error in this problem goes like  $\Delta t^2$  because we use a centered finite difference approximation with error  $\mathcal{O}(\Delta t^2)$  and a geometric mean approximation with error  $\mathcal{O}(\Delta t^2)$ . Use the method of manufactured solutions combined with computing convergence rate to verify the code. Make a test function for checking that the convergence rate is correct.

**e)** Compute the drag force, the gravity force, and the buoyancy force as a function of time. Create a plot with these three forces.

**Hint.** You can either make a function `forces(v, t, plot=None)` that returns the forces (as mesh functions) and `t`, and shows a plot on the screen and also saves the plot to a file with name stored in `plot` if `plot` is not `None`, or you can extend the solver class with computation of forces and include plotting of forces in the visualization class.

f) Compute the velocity of a skydiver in free fall before the parachute opens.

**Hint.** Meade and Struthers [6] provide some data relevant to [skydiving](#). The mass of the human body and equipment can be set to 100 kg. A skydiver in spread-eagle formation has a cross-section of  $0.5 \text{ m}^2$  in the horizontal plane. The density of air decreases varies altitude, but can be taken as constant,  $1 \text{ kg/m}^3$ , for altitudes relevant to skydiving (0-4000 m). The drag coefficient for a man in upright position can be set to 1.2. Start with a zero velocity. A free fall typically has a terminating velocity of 45 m/s. (This value can be used to tune other parameters.)

g) The next task is to simulate a parachute jumper during free fall and after the parachute opens. At time  $t_p$ , the parachute opens and the drag coefficient and the cross-sectional area change dramatically. Use the program to simulate a jump from  $z = 3000 \text{ m}$  to the ground  $z = 0$ . What is the maximum acceleration, measured in units of  $g$ , experienced by the jumper?

**Hint.** Following Meade and Struthers [6], one can set the cross-section area perpendicular to the motion to  $44 \text{ m}^2$  when the parachute is open. Assume that it takes 8 s to increase the area linearly from the original to the final value. The drag coefficient for an open parachute can be taken as 1.8, but tuned using the known value of the typical terminating velocity reached before landing: 5.3 m/s. One can take the drag coefficient as a piecewise constant function with an abrupt change at  $t_p$ . The parachute is typically released after  $t_p = 60 \text{ s}$ , but larger values of  $t_p$  can be used to make plots more illustrative.

Filename: `skydiving.py`.

## Exercise 25: Formulate vertical motion in the atmosphere

Vertical motion of a body in the atmosphere needs to take into account a varying air density if the range of altitudes is many kilometers. In this case,  $\varrho$  varies with the altitude  $z$ . The equation of motion for the body is given in Section 9.8. Let us assume quadratic drag force (otherwise the body has to be very, very small). A differential equation problem for the air density, based on the information for the one-layer atmospheric model in Section 9.6, can be set up as

$$p'(z) = -\frac{Mg}{R^*(T_0 + Lz)}p, \quad (150)$$

$$\varrho = p \frac{M}{R^*T}. \quad (151)$$



To evaluate  $p(z)$  we need the altitude  $z$ . From the principle that the velocity is the derivative of the position we have that

$$z'(t) = v(t), \quad (152)$$

where  $v$  is the velocity of the body.

Explain in detail how the governing equations can be discretized by the Forward Euler and the Crank-Nicolson methods. Discuss pros and cons of the two methods. Filename: `falling_in_variable_density.pdf`.

### Exercise 26: Simulate vertical motion in the atmosphere

Implement the Forward Euler or the Crank-Nicolson scheme derived in Exercise 25. Demonstrate the effect of air density variation on a falling human, e.g., the famous fall of [Felix Baumgartner](#). The drag coefficient can be set to 1.2. Filename: `falling_in_variable_density.py`.

### Exercise 27: Compute $y = |x|$ by solving an ODE

Consider the ODE problem

$$y'(x) = \begin{cases} -1, & x < 0, \\ 1, & x \geq 0 \end{cases} \quad x \in (-1, 1], \quad y(1-) = 1,$$

which has the solution  $y(x) = |x|$ . Using a mesh  $x_0 = -1$ ,  $x_1 = 0$ , and  $x_2 = 1$ , calculate by hand  $y_1$  and  $y_2$  from the Forward Euler, Backward Euler, Crank-Nicolson, and Leapfrog methods. Use all of the former three methods for computing the  $y_1$  value to be used in the Leapfrog calculation of  $y_2$ . Thereafter, visualize how these schemes perform for a uniformly partitioned mesh with  $N = 10$  and  $N = 11$  points. Filename: `signum.py`.

### Exercise 28: Simulate growth of a fortune with random interest rate

The goal of this exercise is to compute the value of a fortune subject to inflation and a random interest rate. Suppose that the inflation is constant at  $i$  percent per year and that the annual interest rate,  $p$ , changes randomly at each time step, starting at some value  $p_0$  at  $t = 0$ . The random change is from a value  $p^n$  at  $t = t_n$  to  $p_n + \Delta p$  with probability 0.25 and  $p_n - \Delta p$  with probability 0.25. No change occurs with probability 0.5. There is also no change if  $p^{n+1}$  exceeds 15 or becomes below 1. Use a time step of one month,  $p_0 = i$ , initial fortune scaled to 1, and simulate 1000 scenarios of length 20 years. Compute the mean evolution of one unit of money and the corresponding standard deviation. Plot the mean curve along with the mean plus one standard deviation and the mean minus one standard deviation. This will illustrate the uncertainty in the mean curve.

**Hint 1.** The following code snippet computes  $p^{n+1}$ :

```
import random

def new_interest_rate(p_n, dp=0.5):
    r = random.random() # uniformly distr. random number in [0,1)
    if 0 <= r < 0.25:
        p_np1 = p_n + dp
    elif 0.25 <= r < 0.5:
        p_np1 = p_n - dp
    else:
        p_np1 = p_n
    return (p_np1 if 1 <= p_np1 <= 15 else p_n)
```

**Hint 2.** If  $u_i(t)$  is the value of the fortune in experiment number  $i$ ,  $i = 0, \dots, N-1$ , the mean evolution of the fortune is

$$\bar{u}(t) = \frac{1}{N} \sum_{i=0}^{N-1} u_i(t),$$

and the standard deviation is

$$s(t) = \sqrt{\frac{1}{N-1} \left( -(\bar{u}(t))^2 + \sum_{i=0}^{N-1} (u_i(t))^2 \right)}.$$

Suppose  $u_i(t)$  is stored in an array `u`. The mean and the standard deviation of the fortune is most efficiently computed by using two accumulation arrays, `sum_u` and `sum_u2`, and performing `sum_u += u` and `sum_u2 += u**2` after every experiment. This technique avoids storing all the  $u_i(t)$  time series for computing the statistics.

Filename: `random_interest.py`.

## Exercise 29: Simulate a population in a changing environment

We shall study a population modeled by (121) where the environment, represented by  $r$  and  $f$ , undergoes changes with time.

**a)** Assume that there is a sudden drop (increase) in the birth (death) rate at time  $t = t_r$ , because of limited nutrition or food supply:

$$a(t) = \begin{cases} r_0, & t < t_r, \\ r_0 - A, & t \geq t_r, \end{cases}$$

This drop in population growth is compensated by a sudden net immigration at time  $t_f > t_r$ :

$$f(t) = \begin{cases} 0, & t < t_f, \\ f_0, & t \geq t_f, \end{cases}$$

Start with  $r_0$  and make  $A > r_0$ . Experiment with these and other parameters to illustrate the interplay of growth and decay in such a problem. Filename: `population_drop.py`.

**b)** Now we assume that the environmental conditions changes periodically with time so that we may take

$$r(t) = r_0 + A \sin\left(\frac{2\pi}{P}t\right).$$

That is, the combined birth and death rate oscillates around  $r_0$  with a maximum change of  $\pm A$  repeating over a period of length  $P$  in time. Set  $f = 0$  and experiment with the other parameters to illustrate typical features of the solution. Filename: `population_osc.py`.

### Exercise 30: Simulate logistic growth

Solve the logistic ODE (122) using a Crank-Nicolson scheme where  $(u^{n+\frac{1}{2}})^2$  is approximated by a *geometric mean*:

$$(u^{n+\frac{1}{2}})^2 \approx u^{n+1}u^n.$$

This trick makes the discrete equation linear in  $u^{n+1}$ . Filename: `logistic_CN.py`.

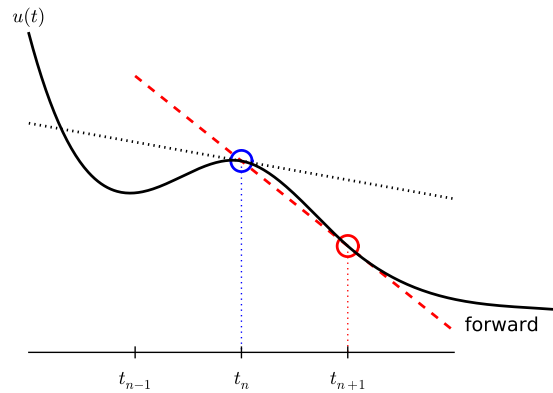
### Exercise 31: Rederive the equation for continuous compound interest

The ODE model (125) was derived under the assumption that  $r$  was constant. Perform an alternative derivation without this assumption: 1) start with (123); 2) introduce a time step  $\Delta t$  instead of  $m$ :  $\Delta t = 1/m$  if  $t$  is measured in years; 3) divide by  $\Delta t$  and take the limit  $\Delta t \rightarrow 0$ . Simulate a case where the inflation is at a constant level  $I$  percent per year and the interest rate oscillates:  $r = -I/2 + r_0 \sin(2\pi t)$ . Compare solutions for  $r_0 = I, 3I/2, 2I$ . Filename: `interest_modeling.py`.

## 11 Summarizing multiple-choice questions

### Exercise 32: Characterize a finite difference

**Question:** We can approximate the derivative at a point using two function values:

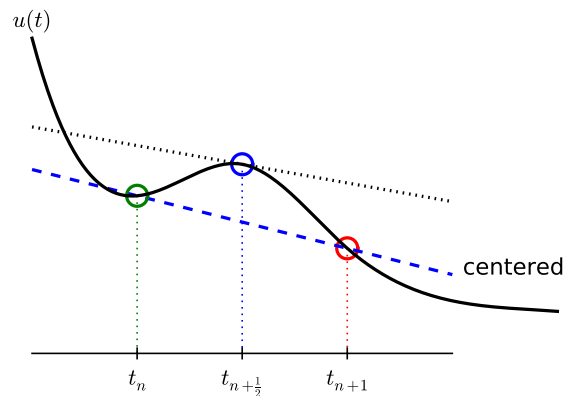


What is this type of difference called and how large is the error?

- A. ☐ This is a centered difference with error proportional to  $\Delta t^2$ .
- B. ☐ This is a forward difference with error proportional to  $\Delta t$ .
- C. ☐ This is a Forward Euler difference with error proportional to  $\Delta t^3$ .
- D. ☐ This is a Backward Euler finite difference with error proportional to  $\Delta t$ .

### Exercise 33: Characterize a finite difference

**Question:** We can approximate the derivative at a point using two function values:



What is this type of difference called and how large is the error?

- A. ☐ This is a centered difference with error proportional to  $\Delta t^2$ .
- B. ☐ This is a forward difference with error proportional to  $\Delta t^2$ .
- C. ☐ This is a centered difference with error proportional to  $\Delta t^4$ .
- D. ☐ This is a Backward Euler finite difference with error proportional to  $\Delta t$ .

### Exercise 34: The $\theta$ rule

**Question:** Why is the  $\theta$ -rule a convenient method? For example,  $u' = -au$  can be solved by

$$\frac{u^{n+1} - u^n}{t_{n+1} - t_n} = -a(\theta u^{n+1} + (1 - \theta)u^n)$$

A. ☐ It is more accurate than the Forward Euler and Backward Euler schemes.

B. ☐ It is easier to implement than, e.g., the Forward and Backward Euler schemes.

C. ☐ It offers the Forward and Backward Euler schemes, the Crank-Nicolson scheme, and more variants in a single formula.

D. ☐ It can be used for other equations than  $u' = -au$ .

### Exercise 35: What is the problem with this program?

**Question:** We want to solve

$$u' = -au, \quad u(0) = 1,$$

by a Forward Euler scheme,

$$u^{n+1} = u^n - a\Delta t u^n,$$

and the following program

```
from numpy import *
u = zeros(10)
u[0] = 1
dt = 0.1
for i in range(10):
    u[i+1] = u[i] - dt*a*u[n]
```

What is the major problem with this program?

A. ☐ `from numpy import *` is not recommended; one should import explicitly the functions needed or do `import numpy as np`.

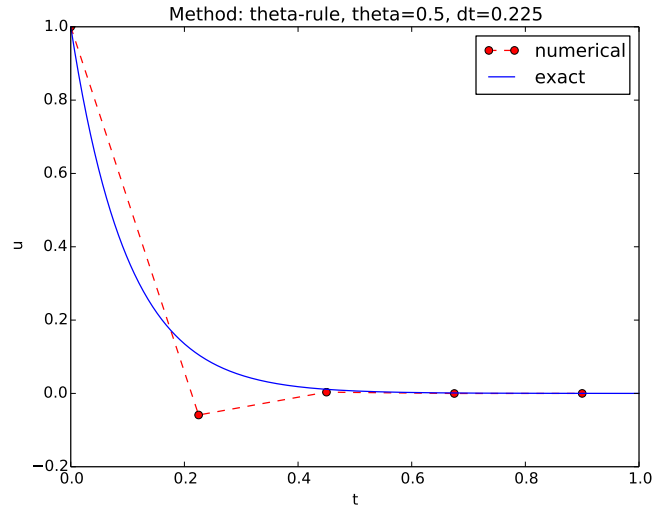
B. ☐ The program aborts with a `NameError`.

C. ☐ The program is “flat”. It should be wrapped in a function `solver(U0, dt, N, a)`.

D. ☐ The scheme is unstable!

### Exercise 36: Is the solution correct?

**Question:** You have solved  $u' = -au$  by the Crank-Nicolson scheme and tested your program thoroughly. Suddenly you run the program with  $I = 1$ ,  $a = 10$ ,  $T = 1$ , and  $\Delta t = 0.225$ . You get somewhat unexpected results:



The results are unexpected because we know the exact solution should be monotone and decreasing, while this numerical solution also shows an increasing stage. What is the problem?

- A. ☐ The program is not tested well enough - yet another bug is there.
- B. ☐ The numerical solution method is more general than the analytical one, and this is an example where the solution can increase, but the analytical solution technique is not capable of dealing with this situation.
- C. ☐ The time step is too large and cause an instability in the form of non-physical oscillations.
- D. ☐ The Crank-Nicolson scheme is unconditionally stable and can be used with all time-step sizes, but the problem here is that round-off errors due to a “too large”  $a$  (10, not around 1) accumulate to the effect seen in the plot.

### Exercise 37: Is this a proper test function?

**Question:** Suppose we have some function `compute` that we want to test. We construct a unit test and implement an associated test function (according to the rules for test functions in the nose or pytest test frameworks):

```
def test_compute(n):
    expected = (n**2)/4.5 - 1
    computed = compute(n)
    assert expected != computed
```

The question is if this test function can be used as is or if improvements must be implemented.

- A. ☐ One should use special assert functions from `nose.tools`, here `nose.tools.assert_almost_equal`.

B. ☐ One cannot test `compute(n)` for only one `n` value. Many are required for good evidence that the function works.

C. ☐ The test function does not test `expected != computed` with a tolerance and the `n` parameter cannot be an argument.

D. ☐ The `assert` statement also needs a message explaining what is wrong when the test fails.

### Exercise 38: Recognize a programming language

**Question:** What kind of programming language is this?

```
function integral = trapezoidal(f, a, b, n)
    % Integrate f from a to b with n intervals
    h = (b-a)/n;
    result = 0.5*f(a) + 0.5*f(b);
    for i = 1:(n-1)
        result = result + f(a + i*h);
    end
    integral = h*result;
end
```

- A. ☐ Python
- B. ☐ MATLAB or Octave
- C. ☐ Cython
- D. ☐ FORTRAN 77

### Exercise 39: Recognize a programming language

**Question:** What kind of programming language is this?

```
def trapezoidal(f, a, b, n):
    # Integrate f from a to b with n intervals
    h = (b-a)/float(n)
    result = 0.5*f(a) + 0.5*f(b)
    for i in range(1, n):
        result += f(a + i*h)
    return h*result
```

- A. ☐ Python
- B. ☐ MATLAB or Octave
- C. ☐ Cython
- D. ☐ FORTRAN 77

### Exercise 40: Recognize a programming language

**Question:** What kind of programming language is this?

```

C      real*8 function trapezoidal(f, a, b, n)
      Integrate f from a to b with n intervals
      real*8 h, result, f
      external f
      h = (b-a)/n
      result = 0.5*f(a) + 0.5*f(b)
      do i = 1, n-1
        result = result + f(a + i*h)
      end do
      trapezoidal = h*result
      return
      end

```

- A. ☐ Python
- B. ☐ MATLAB or Octave
- C. ☐ Cython
- D. ☐ FORTRAN 77

### Exercise 41: Recognize a programming language

**Question:** What kind of programming language is this?

```

double trapezoidal(
  double (*f)(double), double a, double b, int n)
{
  double h, result;
  h = (b-a)/n;
  result = 0.5*f(a) + 0.5*f(b);
  for (i=1; i++; i <= n-1) {
    result += f(a + i*h);
  }
  return h*result;
}

```

- A. ☐ C
- B. ☐ C++
- C. ☐ Cython
- D. ☐ Octave dialect

### Exercise 42: What is SymPy?

**Question:** What is SymPy?

- A. ☐ A Python module for computing with symmetric matrices.
- B. ☐ A Python package for doing symbolic computations (exact/analytical differentiation, integration, equation solves, etc.).
- C. ☐ A Python package for numerical approximations to differentiation, integration, equation solving, etc.
- D. ☐ A free, open source version of Mathematica.



### Exercise 43: What kind of scheme is this?

**Question:** We want to solve  $y' = g(x, y)$  for  $y(x)$  and have the scheme

$$y^{i+1} = y^i + \Delta t g(x_{i+1}, y^{i+1}).$$

What is this scheme called?

- A. ☐ The implicit midpoint scheme.
- B. ☐ The Backward Euler scheme or just the backward scheme.
- C. ☐ The Forward Euler scheme, Euler's method, or just the forward scheme.
- D. ☐ The implicit Adams scheme of order one.

### Exercise 44: What kind of scheme is this?

**Question:** We want to solve  $y' = g(x, y)$  for  $y(x)$  and have the scheme

$$y^{i+1} = y^{i-1} + 2\Delta t g(x_i, y^i).$$

What is this scheme called?

- A. ☐ The implicit midpoint scheme.
- B. ☐ The two-step backward scheme.
- C. ☐ The Crank-Nicolson scheme.
- D. ☐ The leapfrog scheme.

## References

- [1] D. Griffiths, F. David, and D. J. Higham. *Numerical Methods for Ordinary Differential Equations: Initial Value Problems*. Springer, 2010.
- [2] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I. Nonstiff Problems*. Springer, 1993.
- [3] G. Hairer and E. Wanner. *Solving Ordinary Differential Equations II*. Springer, 2010.
- [4] H. P. Langtangen. *A Primer on Scientific Programming With Python*. Texts in Computational Science and Engineering. Springer, fourth edition, 2014.
- [5] H. P. Langtangen and L. Wang. Odespy software package. <https://github.com/hplgit/odespy>.
- [6] D. B. Meade and A. A. Struthers. Differential equations in the new millenium: the parachute problem. *International Journal of Engineering Education*, 15(6):417–424, 1999.
- [7] L. Petzold and U. M. Ascher. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*, volume 61. SIAM, 1998.

## Index

- $\theta$ -rule, 12, 64
- A-stable methods, 46
- Adams-Bashforth scheme, 2nd-order, 66
- Adams-Bashforth scheme, 3rd order, 66
- adaptive time stepping, 71
- algebraic equation, 8
- amplification factor, 45
- array arithmetics, 28
- array computing, 28
- averaging
  - arithmetic, 11
  - geometric, 88
- backward difference, 10
- Backward Euler scheme, 10
- backward scheme, 1-step, 10
- backward scheme, 2-step, 64
- BDF2 scheme, 64
- centered difference, 10
- consistency, 52
- continuous function norms, 28
- convergence, 52
- convergence rate, 61
- Crank-Nicolson scheme, 10
- cropping images, 32
- decay ODE, 5
- difference equation, 8
- directory, 18
- discrete equation, 8
- discrete function norms, 29
- doc strings, 21
- Dormand-Prince Runge-Kutta 4-5 method, 71
- EPS plot, 31
- error
  - amplification factor, 49
  - global, 49
  - norms, 30
- explicit schemes, 64
- exponential decay, 5
- finite difference operator notation, 16
- finite difference scheme, 8
- finite differences, 8
  - backward, 10
  - centered, 10
  - forward, 8
- folder, 18
- format string syntax (Python), 23
- forward difference, 8
- Forward Euler scheme, 8
- geometric mean, 88
- grid, 6
- Heun's method, 65
- implicit schemes, 64
- L-stable methods, 46
- lambda functions, 57
- Leapfrog scheme, 65
- Leapfrog scheme, filtered, 65
- logistic model, 79
- mesh, 6
- mesh function, 6
- mesh function norms, 29
- method of manufactured solutions, 59
- MMS (method of manufactured solutions), 59
- montage program, 32
- norm
  - continuous, 28
  - discrete (mesh function), 29
- ode45, 71
- operator notation, finite differences, 16
- PDF plot, 31
- pdfcrop program, 33
- pdfnup program, 33

- pdftk program, 33
- plotting curves, 24
- PNG plot, 31
- population dynamics, 78
- printf format, 22
  
- radioactive decay, 80
- representative (mesh function), 27
- RK4, 67
- Runge-Kutta, 2nd-order method, 65
- Runge-Kutta, 4th-order method, 67
  
- scalar computing, 30
- scaling, 89
- stability, 45, 52
  
- Taylor-series methods (for ODEs), 66
- terminal velocity, 87
- theta-rule, 12, 64
  
- verification, 62
- viewing graphics files, 31
- visualizing curves, 24
  
- weighted average, 12